Communication, Markets and Society XXXVIII Cycle

A MACHINE LEARNING ENHANCED FRAMEWORK FOR NAVIGATING THE DIGITAL SPACE: RESULTS FROM THE CASE OF FINFLUENCERS

Francesco De Matteo

1039965

Tutor: Prof. Francesco Massara Co-tutor: Prof. Riccardo Manzotti

Coordinator: Prof.ssa Stefania Romenti

Academic Year 2024/2025

INDEX

INTRODUCTION

While web data extraction is not a novelty in academic research, scholars are still very far from exhausting its use. Especially considering its continuous updates and integration with new tools such as machine learning algorithms. This work is inscribed in the web data extraction field and chose finfluencers as the thematic focus to showcase the opportunities that an integrated and updated framework can provide for academic research.

Here multiple objectives are pursued. The first one is relative to the specific thematic domain of choice; finfluencers. A much-needed updated account of the figure of finfluencers is developed in the first part of the document. Financial advice through social media is a topic that evolves way faster than the literature documenting it. As a result, current accounts of the field that have the ambition to be up to date and exhaustive can result already dated. To develop any further research on the topic, which will be done in the following parts of the document, a recent and wide-in-scope account of the figure of finfluencer is both functional and necessary. Second, this work implements a state-of-the-art web data extraction framework and makes the case for wider adoption of this kind of tools. What is shown with this proposal is that machine learning enhanced web data extraction is a productive research and methodological framework that should be more explored from social science scholars and that its recent developments are making its implementation progressively easier and more convenient for non-technically inclined academics. The third aim of this work tackles a specific hypothesis relative to the finfluencer's social media environment. In particular, it is discussed that there is a detectable relation between visual elements that finfluencers use in their social media communication strategies and the thematic context that is contained in the associated text. So much so that it is possible to develop a predictive model that infers thematic context just

from few visual cues in post images. Attributing domain-specific semantical significance to images using minimal visual cues results in two considerable achievements:

- It allows for greatly reducing the time needed for digital space exploration since few visual elements already bring semantic information.
- More importantly it allows to inform areas where text is poor, or absent altogether, of semantic meaning, resulting in the opportunity to infer information where the relevant messages are only implicit; in this way unlocking new domains of data that would have been otherwise unavailable.

Before describing the structure of the document, it is opportune to further define the disciplinary field this work is part of, so to better assess its role in the literature.

Web Data Extraction

Data extraction, or web data extraction, is the process of using automated software systems to obtain structured data from web pages, application programming interfaces (APIs), and other online sources (Cuello, 2024). Historically, scraping involved the process of parsing HTML and retrieving tabular data or listings. Nowadays, it embraces richer processes that gather social posts, pictures, videos, metadata, and conversational threads on a large scale (Brokensha et al., 2019). Extraction at web-scale is foundational to business intelligence, such as price or competitor monitoring, media monitoring, and even social research at large; in the social web space, automated harvesting of public content creates unprecedented opportunities to analyze human behaviors at

enormous scale, empowering researchers to analyze phenomena from public health trackers to urban mobility signals (Ferrara et al., 2014).

In academic settings, data extraction provides resources necessary for computational social science, epidemiology, systematic evidence synthesis, and numerous other applications that require large observational datasets to generate empirical insights (Schmidt et al., 2020). For example, an automated monitoring of foodborne illness reports on Twitter, identification of mental-health indications on Reddit and Twitter, or even extraction of factual results and figures in academic publications to perform meta-analysis (Tao et al., 2023). The advancement of machine learning has not only expanded the range of extractable data but also enabled researchers to transform noisy and multimodal web content into high-quality, structured datasets relevant for statistical analysis and causal inference (Polak & Morgan, 2024).

Extraction pipelines nowadays have multiple steps that transform raw online artifacts into labeled, analyzable records. The first step is data collection through crawling or scraping. This step may use platform APIs or HTML parsers to capture posts, images, comments, and metadata. However, researchers must be careful to balance scale with legal and ethical considerations and API limitations. The second step is preprocessing and natural language processing (NLP), where text is cleaned, tokenized, and then processed using machine learning tools, ranging from classic classifiers to transformer-based models, to extract entities, sentiment, and topic structure (Tao et al., 2023).

The third step is computer vision, where object detection and image understanding models, such as the EfficientDet and YOLO families, convert photos and video frames into object labels, bounding boxes, and scene descriptors that supplement the text features (Wang et al., 2023). The last step is integration or multimodal fusion, where text, vision, and audio outputs are combined through feature concentration, attention-based fusion, or by routing all extracted tokens into a unified reasoning model, such as a multimodal LLM. This last step is increasingly central because when social posts combine captions, pictures, and spoken audio, treating these modalities jointly produces a more coherent representation of the information content of the post (Shetty et al., 2024). Every phase presents specific quality, bias, and provenance issues and benefits from model-driven validation processes to ensure reproducibility (Tao et al., 2023).

Extraction pipelines based on machine learning can have a wide variety of applications. In public health surveillance, social posts have been applied to identify foodborne cases, track disease hot spots, and identify early warning signs of disease outbreaks; dual-task transformer models are also capable of both classifying social posts and extracting entities such as food, symptoms, and location needed for epidemiology follow-up (Salaris et al., 2025). Object detectors used on geotagged images have also been utilized in urban and transportation research to determine bicycle parking demand and road conditions by interpreting user-generated images (Knura et al., 2021). By leveraging both text and image indicators, such as brand mentions in captions and visual logo detections in images, market and brand intelligence approximates reach and sentiment on platforms such as Instagram and TikTok.

For evidence synthesis and systematic reviews, semi-automated methods of extraction, such as BERT+CRF and now, LLM-based workflows, significantly reduce the duration needed to convert unstructured reports into structured trial outcomes and properties (Schmidt et al., 2024). Multimodal fusion of images and text offers further benefits to crisis informatics. Images provide excellent information and visually confirm the extent of damage, while text serves to give location and context. Joint pipelines are much more reliable in determining the extent of harm than unimodal solutions (Shetty et al., 2024). Tao et al. (2023) recognize that social media's inherently multimodal and noisy nature has now become its own form of standardization, where relatively rigid pipelines that incorporate scraping, ML marking, and human evaluation determine its use.

In addition to these primary uses, applied extraction systems can provide various utility-based roles, including scientific applications and organizational value. First, ML extraction allows for near real-time monitoring: automated pipelines use new posts as new data points for an incremental comparative analysis, utilizing incremental refresh metrics, such as frequency of symptom mention, brand visibility, and damage event, for the researcher or service operators to see trends develop in real-time rather than post hoc (Patnaik & Narendra Babu, 2021). In the context of rapid responses, such as outbreak investigations or disaster responses to hazardous events, this is essential because, in some cases, minutes can matter for accuracy and preventing illness (Tao et al., 2023). Second, multimodal extraction allows for expanded construct validity: where text alone is ambiguous, visual evidence or metadata such as geotags and timestamps can help disambiguate intended meaning or location (Radford et al., 2021). For example, a claim of flooding accompanied by geolocated photos is far more informative than text mentions alone (Guo et al., 2025). Third, scalability and reproducibility improve: as soon as ML extractors are validated, they apply the same

criteria across millions of posts, reducing inter-anthropologist variability and enabling longitudinal studies to be designed and conducted reproducibly (Schmidt et al., 2024).

Nonetheless, the pragmatic implementation of these pipelines raises recurring methodological and ethical dilemmas. The main issues involve data quality and bias: whether sampling bias is introduced via platform APIs, diverse access to smartphones, or social media norms unique to the platform, data quality is compromised based on who we see and do not see in our dataset (Nasser et al., 2025). Similarly, object detectors trained on general benchmarks may underperform on the specific images encountered on social media (Polak & Morgan, 2024). As a result, evaluation is paramount; realistic test sets, human-in-the-loop spot checks, and context-specific metrics such as precision for rare event detection and recall for surveillance are required to understand performance trade-offs. Applications are also conditioned by privacy and legality: studies are increasingly drafting minimization strategies, including collecting only public posts, anonymizing data, aggregating, and consulting IRB or platform-specific guidance to mitigate harm. Finally, model transparency and explainability are crucial for practical relevance, especially in public-health or policy-focused applications where decisions made about the extracted data must be defensible to interested stakeholders.

The extraction of textual information from social media and scientific literature has become the primary function of transformer-based language models, including BERT variants, BERTweet, and the GPT family (Huang et al., 2024). While earlier rule-based or keyword-based methods required the laborious process of writing rules and fragile heuristics, pretrained transformers enable robust

entity recognition, relation extraction, and classification with relatively small labeled training sets or through prompt engineering (Shen et al., 2023). Dual-task models, such as BERTweet, have been employed to both recognize the contents of foodborne illness reporting posts, and extracting the related constituents, including food, symptoms, and location, generating significant improvements over a keyword baseline (Hu et al., 2022b). More recently, instruction-tuned LLMs and cloud LLM APIs have been applied to convert free-text pathology and clinical notes into structured fields with extremely high accuracy; evaluations demonstrate that ChatGPT-family models can achieve levels of accuracy that render them potential "second reviewers" or rapid annotation assistants in evidence workflows (Jensen et al., 2025). These models reduce human annotation time and scale entity extraction tasks to corpora that would otherwise be infeasible to process manually. Furthermore, attention is required with hallucination, domain-specific vocabulary, and privacy considerations.

In the case of images and videos, deep object detectors such as EfficientDet, YOLOv5/11, and related families transform pixels into objects labeled and regions of interest (Durve et al., 2023). EfficientDet formalized BiFPN and scaling of compounds to generate efficient detectors, effective for large image corpora (Tan et al., 2021). In contrast, later real-time detectors, for example, YOLOv7, have extended inference speed and accuracy to make it applicable to edge devices and to support the processing of large volumes of users' images (Wang et al., 2023). Applied to social media, these types of vision models detect infrastructure damage, logos, products, or other scene elements features that cannot be conveyed in text form alone, for example, counting the number of bicyclists in the user photos has helped planners estimate informal parking demand (Knura et al., 2021). Combining the outputs of vision with geolocation and timestamp metadata creates

formidable spatio-temporal datasets for urban analytics, catastrophe response, and marketing measures (Tan et al., 2021).

Short-form video platforms and livestreams feature spoken information and onscreen text that are useful for extraction. Automatic speech recognition (ASR), OpenAI's Whisper, processes audio into transcript format, and the transcript is subjected to a series of functions through either a "pipeline" of large language models (LLMs) or classic natural language processing (NLP) pipelines. Optical character recognition (OCR) processes onscreen captioned text (Sharp et al., 2025). When ASR, OCR, image-labeling, and caption parsing components are combined in the processing chain for videos or TikToks, the integration yields stronger, more reliable labels for downstream classification and trend forecasting. Multimodal processing is valuable because the evidence available in audio materials and categorical visual channels generates different but useful forms of evidence. Audio has spontaneous communication behaviors that the captions do not cover, and the ASR recognizes those spoken claims, but the captions do not consider them.

Most actual-world extraction systems are not monolithic single models. They are ensembles or modular pipelines consisting of specialized modules (text classifier, NER, object detector, ASR), which produce features integrated by a downstream classifier or an LLM, reasoning across modalities. Stacking and ensembling are also common in practitioner work: researchers use stacking ensembles of transformer embeddings and traditional learners for mental-health detection (Hridoy et al., 2024). They also employ sequential pipelines where the labels of a vision encoder are employed as additional tokens to decode using an LLM (Ogunleye et al., 2024). Hybrid

architectures such as these leverage the optimal tool for a modality while ensuring engineering flexibility and interpretability.

One of the most trending and rapidly developing approaches is to treat multimodal fusion not as an engineering problem after the fact, but as a first-order modeling objective. Multimodal large language models (MLLMs) that accept both text and images as native inputs and provide text outputs have been developing very quickly, and some of the typical representatives include the LLaVA, GPT-4V, and BLIP families (Zhang et al., 2024b). Benchmarking surveys and studies recognize MLLMs as a key area of research because they acquire cross-modal reasoning, OCR-free chart reading, and end-to-end summarization of image and text data within one pass (Li et al., 2022b). MLLMs can simplify pipelines through the removal of some of the weak glue between vision and language steps, resulting in more coherent, contextualized extraction outputs (Yin et al., 2024). Important caveats continue to hold: computational cost, lack of sensitivity to adversarial or out-of-distribution images, and the necessity of domain adaptation. Despite that, the direction of research and product availability means strong community momentum toward combined vision-language extraction.

Not all research employs an end-to-end MLLM. Numerous high-quality systems utilize specialized, best-in-class modules, for instance, ASR-OCR-object detector-LLM, because the modularity allows these features to be fine-tuned for particular objectives while using inference budgets efficiently and with greater certainty of provenance (Polak & Morgan, 2024). In the case of social media applications, this architecture is reasonable: object detectors can quickly label

thousands of images, Whisper retrieves spoken words with demonstrably low latency, and an LLM provides flexible synthesis and classification. Human-in-the-loop validation and deliberate error analysis, which are essential in high-stakes or regulatory settings, are also easily achieved by the modular approach.

A further benefit of modular pipelines is operational resilience. When one module fails, such as when an object detector misunderstands a new visual style, that module can be retrained or replaced without having to reengineer the whole system (Rosero et al., 2024). Modularity also facilitates cost-performance trade-offs; for instance, lighter computer-vision models could run at the edge for preliminary triage. At the same time, more expensive LLM inference is retained for high-value items, opening the door for staged quality controls, including automated flags and then human review (Ferrag et al., 2025). From a governance perspective, splintered modules offer a better audit trail and provenance capture because each transformation, including ASR transcripts, OCR extracts, and detection labels, is an explicit artifact that can be versioned, tested, and explained. The operationalization of modular orchestration allows replicable research. When the components, interfaces, and explicit data contracts of relevant containerized modules can be shared across teams and domains, sharing and then validating a pipeline is categorically easier, thus enabling deep multimodal reasoning with interpretability.

In the reported uses of the outer-fusion function, modality has been evaluated; evidence consistently demonstrates that fused modalities produce greater downstream extraction and classification accuracy. Cross-modal attention networks and middle-fusion attention networks

allow a model to stabilize visual evidence with textual evidence. This is particularly important for images with somewhat ambiguous or textual captions, especially in relation to crisis scale (Gan et al., 2024). Research has indicated that when the dataset is disaster related and includes images and text, for classification of relevance accuracy, the relevant accuracy collectively improved by an average of between 2% and 5% (Shetty et al., 2024). From a large-scale perspective, a significant improvement for the detection of rare explicit events or enhanced sensitivity for public health or population health surveillance, and particularly 'thin signals'.

The rapid growth of MLLMs has produced multiple overview and method articles establishing the research agenda that includes levels of vision-language comprehension, reliable evaluation practices, and prompt engineering and instruction tuning best practices (Yin, 2024). Public models and open toolkit models such as LLaVA, BLIP/BLIP-2, and CLIP continue to improve reproducibility, while extending multimodal experimentation beyond larger-scale research labs (Li et al., 2023). At the same time, scientific evidence has accumulated regarding the semi-automation of systematic reviews, ethical practices for scraping, and judgments on legal responsibilities. These are helping to ground model-driven developments in a more pragmatic direction towards research realities in practical contexts (Mutlu et al., 2024).

The research, along with the toolkits and benchmarks, is now also progressing toward reproducible pipelines, evaluations in realistic distributional shifts, and governance-ready practices, creating an entirely new space between an experimental prototype and a deployable system. Some illustrative examples include existing datasets with a satisfactory level of documentation, presence of

provenance about the data and data-sheets; standardized evaluation suites designed to test cross-modal robustness; and adversarial or rare-event situations designed to probe potential brittleness of a model (Cannata et al, 2023). From a prioritization perspective, methods such as domain adaptation, few-shot learning, active-learning methods that address annotation costs, and synthetic-data augmentation help to improve the understanding of scenarios likely underrepresented in social media images and vernacular text (Li et al., 2022). From an engineering point of view, event-stream and MLOps/orchestration frameworks are required to help ensure repeatability in experiments and safety in rollouts, such as containerized components, CI/CD processes for models, and version control on artifacts (Eken et al, 2025).

Ethically, there has been a convergence of best practices around privacy-preserving collection, differential minimization, and transparent reporting, including model cards and impact statements, so that extraction works are auditable and socially accountable. Community-driven shared tasks, benchmark competitions, and interdisciplinary workshops are fast-tracking consensus on evaluation practices or data standards, and producing an ecosystem where methodological innovations or advances can be rigorously compared, responsively scaled, and accelerated to operational research and public-interest applications (El-Moussaoui et al., 2025).

Data extraction has transitioned from rule-based HTML parsing into a robust, model-based discipline in which machine learning interprets, labels, and synthesizes multimodal web content at scale. In academia and practice, in more practical uses, particularly in public health, urban planning, brand analytics, and even evidence synthesis, ML-based extraction pipelines can lessen

the amount of manual work and be used to offer previously impractical analyses. The recent development of the most interest is the emergence of multimodal approaches that coordinate models across modalities intentionally or adopt unified MLLMs. These combined designs are more suitable for the multimodal aspect of social media posts and are more likely to produce valuable and accurate extracted datasets consistently. The practical course today takes a data-collector or practitioner to mix by the rigorous approach to data collection and responsible practice with the components that are separably trained by ML models - and, increasingly, it challenges whether a multimodal LLM can simplify procedures or increase precision. This multimodal, integrated approach is not just trendy; it is a meaningful technical breakthrough, widening the range of what can be found using the social web.

Structure of the Work

The following section will present a wide-scoped desk review on the topic of finfluencers. The necessity for keeping the disciplinary field of reference as wide as possible derives from the fact that academic literature alone is not sufficient to return an exhaustive view of finfluencers as a figure, since their role and characteristics are heavily dependent on both followers/investors' behavior and the current stance of major regulatory authorities. The section will delve into definitional boundaries and scope, distinguishing education-first from trading-first content and clarifying jurisdictional limits. Historical emergence is traced to low-friction brokerage, short-video formats, and pandemic-era dynamics. Communication strategies are examined through strategic authenticity, parasocial bonding, and platform-native rhetoric that convert attention into action. Topic clusters and monetization are mapped, spanning literacy content and memberships as well as trading and crypto funnels with affiliate rewards. Ethical and professional risks are

reviewed, including conflicts of interest, suitability gaps, vulnerable audiences, limits on truthfulness and balance, and manipulation. The regulatory landscape is summarized across major jurisdictions with converging expectations on disclosure, fairness, and prior approvals. Effectiveness is assessed across market impact, investor welfare, and educational benefit, with heterogeneity and selection noted. Platform and algorithmic dynamics are analyzed for their role in exposure, disclosure salience, and diffusion. Remaining research gaps are identified in long-run outcomes, algorithmic causality, non-English ecosystems, and creator economics. Methods underpinning the review are outlined, including eligibility, search, screening, appraisal, bias controls, ethics, and limitations. Results are synthesized on identity, formats, engagement devices, and outcomes. Stakeholder recommendations are set out for regulators, platforms, firms, creators, educators, and researchers, followed by monitoring metrics and a staged implementation roadmap.

The following section is instead dedicated to introducing and describing the technical tools that will be used to build, enrich and analyze the dataset that will be used in the last section experiment. Namely web crawling, LLMs models and object detection. For each one of the topics it will be outlined a brief history and how it has been used in academic research. In the specific case of web crawling and scraping a more extensive part will be dedicated to the ethical implications of the use of this kind of instrument and the regulatory conditions that discipline their implementation; with the aim of identifying an ethically and regulatory compliant implementation.

The penultimate section will implement an LLM-powered thematic analysis of Instagram post captions, to both develop an updated account of the discourse that finfluencers conduct on social media and arguing for the use of LLMs in the processing of extensive volumes of unstructured

textual data. In particular, the analysis will make use of the *Instagram Finfluencers Database*; a dataset of 22.854 public captions from 01-2024 to 09-2025, collected via a custom scraper, then apply reflexive thematic analysis assisted by an LLM. The study addresses three questions: which topics are the most discussed in the captions, how these topics express empowerment narratives, and whether an LLM can support exhaustive TA at scale. Results will output 36 main themes that connotate the content of recent finfluencers social media activity.

In the last section a predictive model funded on random forest analysis will be developed. The model will be trained and tested with a K-fold cross-validation approach on an enriched dataset derived from the one already used for the previously executed thematic analysis. The dataset, enriched by information about visual cues in the images of the posts via the use of YOLOv11 object detection model and by the remapping of themes to single captions, will be used to validate the effectiveness of exploiting visual cues as features with the purpose of predicting the thematic content of the captions. Presentation and discussion of empirical results will follow; while two appendixes at the end of the document, Appendix A and B, will store respectively the summary table of the thematic analysis and the numerical results of the random forest analysis.

FINFLUENCERS

Background and Context

Money advice has moved into everyday feeds. In a few years, millions of people began to learn about budgeting, investing, and crypto from creators they follow on YouTube, TikTok, Instagram, Reddit, and X. The popular label for these creators is finfluencers. The term describes people whose main output is financial content and whose voice blends education, entertainment, and promotion (Hasanah et al., 2025). The attraction is easy to see. Their videos feel informal and direct. Their posts use plain language and concrete steps. They share mistakes and small victories, which contribute to the safety of the topic in the conversation. Both surveys and professional reports indicate that a significant number of young adults worldwide now learn about money concepts and solutions first through creator content instead of at school or through licensed advisers and that they experience this channel as being more relatable and more transparent about lived experience than standard finance media (Pokhrel et al., 2025). The development of this focus has changed the public space of financial information and raised worrisome questions related to research and policy.

The background to this change is technological and social. App-based brokerage, fractional shares, instant account funding, and zero commission trading lowered the cost and effort required to act on a new idea. Behavioral finance work documents how mobile interfaces and push notifications can nudge frequent trading and shorten holding periods, which pulls people toward event-driven decisions that are highly responsive to salient stories and social signals (Barber et al., 2022). Meanwhile, the advent of short video clips and live streaming made producing content that looks and sounds credible inexpensive. There are countless formats designed to display catchy hooks, numbered lists, and phrases of high certainty, while algorithmic feeds amplify messages that sound

new and certain at the expense of slow, reasoned explanations (Shin, 2022; De Veirman et al., 2017; Vosoughi et al., 2018; Bischoff et al., 2019). The pandemic added time at home, market volatility, and a desire for community, which helped retail investors form dense online networks that reward participation and quick feedback (Shiller, 2019; Cookson et al., 2023; Ante, 2023). Unsurprisingly, creators who speak clearly about money grew fast in this environment.

There is real potential in the global expansion of finfluencers. They reduce the barriers to entry to financial learning. Marketing and media research have identified two characteristics that account for a significant portion of their popularity (Abidin, 2018; Pittman & Abell, 2021; Hugh Wilkie et al., 2022). The first is strategic authenticity. Creators stage ordinary settings, share personal narratives, and maintain a consistent value frame that signals they are like the audience and not above them. The second is the parasocial bond. By continually bombarding the target market with a common face and voice, a level of familiarity and trust is created that reduces their skepticism and encourages them to act (Harmeling et al., 2017). Both effects are quantifiable, and they both forecast intention to use a recommended product or adhere to a proposed plan, particularly when the person behind them feels similar to the recipient (Sokolova & Kefi, 2020; Lou & Yuan, 2019; Reinikainen et al., 2020; Hii & Ong, 2025). Recent analysis centering on finance indicates that authenticity is a constructed product and that creators are learning the grammar of every format to optimize their spread, whilst maintaining an informal tone of peer-to-peer advice (Zhu et al., 2019).

The same conditions that make finfluencers effective also produce new forms of risk. Most creator messages are broadcast to large, heterogeneous audiences rather than tailored to a single client with documented goals and risk capacity (Casaló et al., 2020). Consequently, content that appears to be

general education may be perceived as personal advice by followers with a strong parasocial bond. In the case of a risky product, the result may be detrimental even when the creator did not mean to deceive. Finance research has found that social attention can shift prices in the short term and can promote lottery-like positions and short holding periods that lower subsequent returns to the investors who respond most to the signal (Cookson et al., 2023; Barber et al., 2022; Merkley et al., 2024; Warkulat & Pelster, 2024). This trend does not imply that finfluencers are necessarily bad. Scale, speed, and human psychological behavior can combine to circumvent the defense mechanisms surrounding traditional one-to-one advice.

The Regulatory Landscape

The regulatory landscape reflects these tensions. Supervisors in several jurisdictions have adapted long-standing principles to social channels and clarified that many creators' posts count as financial promotions or investment recommendations. In the United Kingdom, any message that encourages or persuades to interact with a product must be fair, clear, and not misleading, and most promotions must be approved by a firm even when the sender is an individual creator (ESMA, 2024). In the European Union, transparency requirements can be triggered when an individual makes a public statement that can influence investment decisions under the Market Abuse Regulation (ESMA, 2024; Pokhrel et at., 2025). In the United States, the investment adviser marketing rule sets the conditions of using testimonials and performance claims, and consumer protection authorities require clear and conspicuous disclosure of material connections in any endorsement (Xu & Pratt, 2018). In Australia, there is a practical difference between factual information, which is permitted, and financial product advice, which requires a license, even when the medium is an informal post or video (Boerman et al., 2017). India has moved to limit commercial relationships between

regulated intermediaries and unregistered creators in light of repeated cases where paid promotions were masked as neutral education. International bodies report coordinated actions against illegal promotions and manipulation schemes that use creator channels to seed demand in thin markets (Aral & Eckles, 2019). These frameworks point toward the same core principle. People should be able to tell when a message is sponsored, and they should receive a fair presentation of risks and benefits regardless of the channel.

The rationale for the present study emerges from the changing role of finfluencers in the contemporary financial landscape. Once regarded as a novelty, they have become a stable and influential feature of the information environment. Despite their prominence, the academic discourse remains fragmented across fields and jurisdictions, creating an incomplete picture of their overall impact. Finance and accounting studies primarily concentrate on short-term price fluctuations and attention spillovers. However, these contributions provide limited evidence regarding the long-term effects on household decision-making and outcomes (Sahni, 2016; Jantan, 2024). Communication and marketing research have been valuable in explaining mechanisms of persuasion, though such work often positions finance as a conventional product category rather than a domain shaped by fiduciary duties and market integrity. Policy reports have advanced broad principles for regulation, but they continue to face challenges in evaluating compliance and assessing consumer understanding on a meaningful scale. These limitations point to the necessity of synthesizing evidence across disciplines and designing research that identifies what truly benefits or harms followers over time (Hayes & Ben-Shmuel, 2024; OECD, 2021; Gerritsen & Regt, 2025).

The Significance of the Study

The significance of the study is practical as well as scholarly. For households, creator content can open a door to simple, low-cost strategies or pull a novice into complex products with high spreads and volatility. For firms, partnerships with creators can widen access to tools that serve people well or propagate weak disclosures and unrealistic claims. For platforms, evidence can guide product decisions about labels, prompts, and risk presentation, improving comprehension without suppressing speech. Each of these actors makes design choices. The closer we can map outcomes to choices, the more useful the guidance will be (Hammer, 2025).

This study, therefore, begins with a clear problem statement. The growth of finfluencer content has outpaced the frameworks that ensure fairness, transparency, and suitability for audiences who often treat broadcast messages as personal advice. There is mounting evidence that attention can move markets in the short term and that investors who follow social signals concentrate in high-variance positions with lower subsequent returns.

Treating these topics also demands a serious discussion about the ethical framework in which this study collocates in respect of both financial creators as well as researchers that intend to study them. Pertaining to creators, some core ethical principles such as autonomy (respecting users' ability to understand and consent), non-maleficence (avoiding intentional financial harm), fairness and distributive justice (how harm can be concentrated on vulnerable and financially stressed groups), and accountability and responsibility (for creators, firms, platforms, and regulators) will be discussed and advocated for. In addition, it is worth noticing that some of those principles are directly applicable to the practices of researchers and their instrumental use of AI. In fact, this study

will also engage in building a framework for AI use in research that conforms with principles of transparency, fairness and accountability.

Key Research Questions and the Structure of the Dissertation

The study is guided by three theoretical lenses that help connect micro-level persuasion to macro-level outcomes. The first lens is the influencer credibility and authenticity literature, which explains how perceived expertise, trustworthiness, and congruence between the creator and the product drive persuasion and referral, and how authenticity can be performed without being deceptive when values and messages remain consistent (Sokolova & Kefi, 2020; Lou & Yuan, 2019; Belanche et al., 2021). The second lens is parasocial interaction and community theory, which explains how routine exposure to an approachable figure reduces psychological distance and how community rituals amplify confidence and action, especially for complex or intimidating topics (Reinikainen et al., 2020; Hii & Ong, 2025; Gass & Seiter, 2022). The third lens is platform and algorithm research, which looks at how ranking, watch time, and novelty rewards influence the distribution of messages and how quickly some frames spread through a population. This lens is essential since it places the power in the hands of the messenger and the design of the system in regard to what the audience will see (Shin, 2022; Vosoughi et al., 2018; Gass & Seiter, 2022).

Derived from these lenses, the following research questions guide the dissertation. First, how do scholars and regulatory authorities use the and define the term to finfluencers, and how do these organizations distinguish between content that delivers a more educational approach and content that delivers a trading-oriented approach? Second, what sociocultural trends and technological innovations have driven the high emergence of influencers in financial markets over the last 10

years? Third, what communication patterns lie behind successful finfluencers, and how do the concepts of authenticity, performance metrics, and parasocial connections influence the creation and consumption of financial content? Fourth, what are some of the most significant ethical and legal issues arising from delivering broadcast investment advice services and the commercial incentives of registering users and executing trades? The discussion is organized around three main research questions. First, how effective is finfluencer advice on household financial outcomes in the short and long term? Second, what have regulators' responses been in the major jurisdictions, and how do the regulatory frameworks converge or diverge in ways critical to consumer protection and market integrity? Third, what are the remaining gaps in the literature, especially related to algorithmic causality and cross-platform exposure, and how can future research better directly evaluate the results related to sustained creator contact?

These questions shape the general outline of the dissertation. The next chapter offers an extensive literature review summarizing the available research on finfluencers in finance, marketing, media, and policy, and suggests a conceptual framework based on access, appeal, and accountability. The following one specifies the methodological framework, describes the relevant data sources, and rationalizes the analytic decisions to explore content, disclosures, and outcomes across platforms and jurisdictions. Then a section presents empirical findings on communication strategies and disclosure quality using coded samples of creator posts, while the following one links exposure to outcomes by testing how changes in disclosure design and platform prompts affect comprehension and stated intention among novice investors, and by reviewing observational evidence where appropriate. The second-last chapter of the section compares regulatory regimes, maps responsibilities for firms and creators, and discusses enforcement styles and education initiatives.

The last one synthesizes contributions, discusses limitations, and outlines future research on long-run outcomes, creator economics, and platform design.

A brief note on definitions and scope helps orient the reader. The term finfluencer is used in a functional sense. It is used to describe creators on large platforms whose main content is related to personal finance, investing, trading, or crypto. Its messages may educate, recommend, or promote financial behaviors or products to broad audiences. The term implies nothing about whether the person is licensed; however, most creators are not certified. The research is limited to the United Kingdom and the European Union, the United States, Australia, and India, as these are the jurisdictions that have generated the most uniform set of responses, and some of the most trenchant criticism. Based on the experience of other jurisdictions having an active supervision process and public campaigns (ESMA, 2024), where possible, the review identifies the lessons learned in this area in the respective jurisdiction.

The contributions of the dissertation are threefold. First, it integrates scattered findings across fields into a framework that connects persuasion mechanics to measurable outcomes in retail finance. Second, it offers evidence on disclosure quality and comprehension in creator content, informed by the latest regulatory guidance on fair balance and material connections. Third, it charts regulatory convergence and divergence in a manner that assists practitioners in designing compliance programs in creator partnerships without sacrificing the human elements that make good education effective. These contributions are intended to assist regulators in tuning regulations, companies in exploring safer business practices, and content creators in ensuring trustworthiness

and adopting business approaches that will not harm other people because of the technology creator's negligence (Thorson et al., 2021).

The fact that such scale and intimacy exist in creator culture leads to the need for careful management. One post can reach millions at negligible cost. One creator can cultivate a sense of friendship with thousands of viewers who know the person on screen. That intimacy is a strength when the message concerns budgeting, fee awareness, and patient diversification (Dubois et al., 2016). It becomes a liability when the message invites rapid action in a thin market or a sponsored script hides conflicts of interest. A policy set up to work on television or printed paper does not necessarily apply to a mobile feed refreshed every several seconds. The difficulty is that to use these old principles and make them applicable to new channels and to test disclosure designs and platform prompts to alter understanding and behavior of these media (ESMA, 2024; Kim et al., 2015).

Definitional Boundaries and Scope

Across the last decade, scholars and supervisors converged on a functional definition of the finfluencer. According to the European Securities and Markets Authority (ESMA, 2024), the term is not primarily used as a professional designation but rather refers to content creators whose primary production relates to financial issues and markets. Their posts can educate, advise, or shape the economic behaviors of audiences with whom they are not personally acquainted. This conceptualization as effect lacks technical fondness. The framework spreads easily through short formats, livestreams, long tutorial articles, thread posts, newsletters, and exclusive communities, and is thought to focus on results instead of label titles (ESMA, 2024). Within that frame, the

literature usually distinguishes education first channels from trading first channels. Hammer (2025) explained that education-first accounts emphasize budgeting, debt relief, fee awareness, and diversified investing. In contrast, trading-first accounts foreground instruments, timing, leverage, and frequent repositioning, often illustrated with screenshots of positions or realized gains. The distinction matters because it maps to different legal triggers and audience outcomes. When posts invite or induce engagement with specific products, they cross into promotion and recommendation regimes even if the messenger is not licensed (ESMA, 2024; Boerman et al., 2017).

Hudders et al. (2021), Appel et al. (2020), and Martinez-Lopez et al. (2020) treated finfluencers as a socio-technical construct, emphasizing that the creator is shaped not only by the message but also by the platform grammar through which the message is perceived and by commercial incentives that influence the motivation to speak. Such framing facilitates the comparison of such content between platforms, so as not to move the conceptual goalposts each time a new feature alters all the formats. It also does not create a false dichotomy between education and advertising as the same channel often merges the two in a content calendar (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; Stubb et al., 2019; Wojdynski, 2016).

Historical Emergence in Context

Graf-Vlachy et al. (2018) and Hasanah et al. (2025) indicated that the finfluencer phenomenon reflects broader technological, media, and social behavior shifts. Technological developments such as app-based brokerage, instant funding, and fractional shares have lowered the cost of action for small investors. Behavioral finance work shows that mobile interfaces, push alerts, and frictionless execution nudge shorter holding periods and event-driven trades in which social cues loom large (Barber et al., 2022; Cookson et al., 2023; Belanche et al., 2021). On the media front, short video

and live streaming enable the packaging of financial concepts in a form that feels comfortable to discuss. WHO recommender studies note that novelty, clear messaging, and visible confidence are all rewarded more highly than circumspection, and that probably explains why bold, step-by-step scripts get further than the cautious descriptions (Shin, 2022; De Veirman et al., 2017; Vosoughi et al., 2018; Appel et al., 2020; WHO. 2021). The social effects of the pandemic included increased time availability, heightened volatility, and a stronger desire for shared experiences. Retail participation surged and clustered in online communities where feedback is quick and identity is performed through shared rituals and language (WHO, 2021; Shiller, 2019; Ante, 2023; Cookson et al., 2023; Merkley et al., 2024).

Cultural economy scholars emphasize that the rise of finfluencers is also a story about the normalization of money talk. Abidin (2016) traced how creators bring private financial concerns into intimate spaces and narrate money as a journey of resilience and aspiration. Hayes & Ben-Shmuel (2024) situates finfluencers within a broader moral economy in which thrift, hustle, and self-investment are cast as virtues that align personal identity with market participation. Policy surveys from Europe and North America show that many young adults now encounter financial concepts first through creator content rather than in school or with licensed professionals, and they cite relatability and plain language as reasons they return to the same channels; these threads position finfluencers as mediators who translate market complexity into everyday routines (Zhang & Gong, 2021).

Evolution of the Figure and Market Structure

Two developments characterize the post-pandemic years. The first is professionalization. Many creators now operate as micro media firms with diversified income streams that include sponsorships, affiliate programs, creator fund payments, premium communities, courses, and live events. According to Mölders et al. (2025), authenticity in this setting is not a static trait but a staged performance sustained through scripts, settings, pacing, and visual choices that maintain a peer-to-peer tone even when production is professional. Zhu et al. (2019) similarly show that creators learn platform-specific grammars for being "real" and that this learning differentiates durable channels from transient bursts of virality. There is also a regulatory convergence. In the United Kingdom, the European Union, Australia, India, Singapore, and the United States social posts can constitute financial promotions or recommendations in investments, and as such, supervisors have made explicit that such social posts would be subject to disclosure, fair balance and recordkeeping requirements (ESMA, 2024). IOSCO's final report highlights coordinated actions on illegal promotions and market abuse executed through creator channels (ESMA, 2024). The upshot is that influencers, firms, and platforms operate under increasingly clear expectations, even as enforcement styles differ.

A third element of the evolution is platform ecology. Short-form video favors crisp benefits, visible social proof, and demonstrations that compress action into a few taps. YouTube sustains longer arcs and deeper tutorials that build perceived expertise over time. X and Reddit reward immediacy, debate, and the community co-production of market commentary, whereas Instagram and Facebook embed money themes in lifestyle imagery. Comparative work suggests that creators who align pacing, tone, and calls to action with each platform's recommendation logic receive stronger

algorithmic amplification and thus accumulate more out-of-segment viewers who are especially responsive to confident frames (Shin, 2022; De Veirman et al., 2017; Zhu et al., 2019; Guess et al., 2019).

Communication Strategies: How Finfluencers Persuade

The literature converges on a persuasion stack with three layers: strategic authenticity, parasocial bonding, and platform-native rhetoric. These layers interact to lower psychological defenses and create simple paths from knowledge to action.

Strategic Authenticity

Authenticity is performed through ordinary settings, plain speech, and selective self-disclosure. According to Mölders et al. (2025), finance creators use carefully calibrated pacing and edits to preserve an everyday vibe while delivering scripted and tested lessons for impact. Sokolova and Kefi (2020) show that perceived authenticity and credibility predict intention to take advice across influencer contexts. Lou and Yuan (2019) find that message value and source congruence build trust in branded content. Zhu et al. (2019) add that "acting real" is a skill that creators learn and refine, and audiences reward consistency over time more than they reward one-time disclosures. In finance, authenticity is often anchored by visible spreadsheets, on-screen calculators, and screenshots that stand in for proof, which makes abstract ideas more concrete.

Parasocial Relationships and Community Rituals

Hii & Ong (2025) discussed that parasocial bonding in financial contexts predicts willingness to join paid communities and to follow referral links, with effects moderated by perceived expertise

and homophily. Reinikainen et al. (2020) show that ritual openings, direct address, and comment replies strengthen one-sided bonds that feel reciprocal. Kay et al. (2020) find that micro creators with tighter communities can outperform larger accounts in engagement because closeness raises trust. This mechanism is especially important for personal finance, where embarrassment can suppress questions in formal settings. By presenting money talk as an ongoing conversation with a familiar face, creators make the topic less threatening and more actionable (Canatan et al., 2023).

Platform-Native Rhetoric and the Logic of the Feed

According to the European Securities and Markets Authority (ESMA, 2024), platform-native rhetoric within short feeds rewards hooks, lists, and clear steps. While many videos include disclaimers such as "this is not financial advice," ESMA (2024) emphasized that policy direction cautions against treating boilerplate statements as substitutes for the transparent disclosure of significant connections, which may be unnoticeable to the casual observer. As empirical studies show, smaller or fast-paced labels go unnoticed; on the other hand, a clear disclosure, placed on the first page of the advertisement, increases ad recognition and reduces perceived deception but has relatively weak impacts on user activities (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; Hudders et al., 2021; Stubb et al., 2019). Finfluencers who understand this reality adopt front-of-message disclosures and balance promised benefits with visible risks when they intend to keep trust over time.

Topic Clusters and the Commercial Funnel

Pokhrel et al. (2025) identified five recurring clusters that appear across markets. The first covers literacy basics such as budgeting, emergency funds, and credit building. The second covers diversified long-term investing and retirement planning. The third promotes active trading of

equities and options. The fourth focuses on crypto assets and adjacent products. The fifth ties money to lifestyle through side income, travel points, and negotiation scripts (Pokhrel et al., 2025). Monetization maps onto these clusters. Education-first channels monetize through books, courses, worksheets, and memberships. In contrast, trading and crypto channels often monetize through affiliate links and bounties that pay on sign-ups or trades. The mapping is central to ethical analysis because revenue design shapes what is emphasized and how strongly it is pitched (Beichert et al., 2024).

Ethical and Professional Tensions

Since the work of financial influencers touches on the intersection of education, entertainment, and promotion, issues related to conflicts of interest, appropriateness, protecting vulnerable audiences, honesty, and market honesty may be triggered (Casaló et al., 2020; Anwar et al., 2024).

Conflicts of Interest and Transparency

Rachmad (2024) and Barta et al. (2023) highlighted that one of the most problematic issues is non-disclosed or inappropriately disclosed payments. In the United States, the endorsement guides recommend clearly and consistently disclosing material connections such as affiliate payments and sponsorships. At the same time, the adviser marketing rule restricts testimonials and performance claims by registered firms. In the United Kingdom, posts that tempt or encourage the person to come into contact with products are promotions that should be fair, precise, and not misleading, and most of them require the firm's consent (ESMA, 2024). In the European Union, consultations with the market that may compel investment decisions may give rise to obligations associated with the Market Abuse Regulation (ESMA, 2024; Armour, 2021). This is because Australia divides

between objective data and guidance that directs a choice, which can be licensed even as a casual video (Mason & Clarke, 2025). India opted to regulate the commercial dynamics between regulated intermediaries and unregistered creators after several instances of paid promotions that were framed as objective informational material despite being advertisements (Singh, 2025). The unifying principle of these frameworks is straightforward. The audience must understand that it is an advertisement and should get an equal representation of risks and benefits (Huang et al., 2020).

Suitability and the Broadcast Problem

Hii and Ong (2025) noted that licensed advisers are required to know their customers and to document suitability. Creator messages are broadcast to heterogeneous audiences. The unavoidable result is a suitability gap that no disclaimer can fully close. Survey work shows that followers often treat general guidance as personal advice when the messenger appears relatable and credible, which raises the risk of misfit decisions in complex products (Hii & Ong, 2025; Abidin, 2018; Khoirotunnisa, 2024). Supervisors warn firms that approve creator content to avoid personalization cues unless the relationship clearly triggers advice duties and to ensure that risk information is as salient as benefit claims in high-risk content (ESMA, 2024).

Vulnerable Audiences and Distributive Effects

Barber et al. (2022), Warkulat and Pelster (2024), and Merkley et al. (2024) provided evidence that social attention is linked to higher risk-taking and shorter holding periods, which in turn reduce subsequent returns for investors who pursue excitement, particularly in thin assets and leveraged instruments. Young and financially stressed audiences are overrepresented among followers of trading first channels, which concentrates the harm. In comparison, education-first channels

promote fee knowledge and low-cost diversification in line with core professional recommendations (Pokhrel et al., 2025). These facts show that the results depend not only on the content style but also on the audience structure. The fact that user exploitability and potential harm is concentrated results as particularly problematic for ethical reasons. In particular, an audience structure in which young and financially stressed users are the most vulnerable segments also makes principles of fairness and autonomy easy to violate implicitly for financial creators.

Truthfulness, Balance, and Format Constraints

Hammer (2025) noted that consumer law requires truthfulness and prohibits misleading omissions and exaggeration. Financial promotions add a duty to balance risks and benefits and to avoid cherry-picking performance. Short formats and algorithmic incentives make balance harder. Regulatory bodies have reacted with visible examples of adequate risk disclosure and have clarified that disclosure cannot be hidden in captions or be made via links to outside websites (Hammer, 2025). The voluntary initiatives, like the French certificate of the responsible influence in finance, also seek to pick up the lowest ground by increasing mutual literacy between the creators and the brand (Graf-Vlachy et al., 2018; Schivinski & Dabrowski, 2016).

Market Integrity and Manipulation

Ante (2023), Cookson et al. (2023), and Merkley et al. (2024) found that large accounts can move thin markets. Studies of crypto and meme equities show immediate spikes in price and trading activity around prominent posts. These effects are typically followed by reversals, which the authors interpret as attention-driven overreaction rather than the incorporation of durable information. Supervisors report coordinated actions against illegal promotions, scalping, and pump

and dump schemes that use creator reach to seed demand (Aral & Eckles, 2019). The policy lesson is that digital channels change the speed and reach of old problems, not the underlying logic.

Effectiveness and "Performance" Of Finfluencer Advice

Effectiveness depends on the outcome of interest and the horizon of analysis. The literature separates market impact, investor welfare, and educational benefit, stressing heterogeneity across creator types.

Market Impact

Ante (2023) and Merkley et al. (2024) showed through event-style studies that creator posts can quickly influence prices and trading volume. Their analyses indicate that prominent tweets and videos in crypto correlate with sharp market reactions that fade within days. A pattern consistent with attention-driven liquidity effects rather than persistent alpha. Retail attention predicts next-day returns in equities linked to online communities, followed by reversal (Cookson et al., 2023). These findings do not condemn creator content; they show that attention is a tradable force that can detach prices from fundamentals in the short term.

Investor Welfare

When analysis shifts from markets to people, results are more cautious. Using trading data, Barber et al. (2022) show that bursts of social attention increase the probability that retail investors take lottery-like positions, shorten holding periods, and underperform diversified benchmarks after risk adjustment. Warkulat & Pelster (2024) links social attention to concentrated, high-variance bets

with low subsequent returns. The mechanism resembles a mix of salience bias and social proof that creator communities can amplify.

Educational Benefit

Pokhrel et al. (2025) found that education-first content shows promise. Surveys and experiments indicate that confidence and the uptake of low-cost strategies can be improved through exposure to creators who stress budgeting, fee awareness, and diversified long-run investing, particularly among new investors without access to professional planning (Pokhrel et al., 2025).

Heterogeneity and Selection

Warkulat and Pelster (2024) observed that the label *influencer* conceals wide variation, as some creators are licensed and operate within compliance, whereas others promote high-pressure funnels into leveraged products. They further noted that selection complicates causal claims, since audiences who seek risky content may already prefer variance. Recent work uses exogenous shocks to attention and policy changes to improve identification, but the literature still calls for long-horizon panels that link exposure and outcomes over time (Warkulat & Pelster, 2024; Merkley et al., 2024; Hudders et al., 2021). To evaluate "performance," several authors propose a practical benchmark: outcomes relative to a low-cost diversified strategy net of fees. Education-first content aligns with this benchmark for most households, while trading-first scripts move followers away from it and increase timing risk and costs (Foerster et al., 2017; Barber et al., 2022).

How Academia Has Treated Finfluencers So Far

The research arc runs from description to measurement to policy. Initial media and cultural economy commentators referred to creators as financializing agents integrating market engagement into daily life by narrating intimate stories and aspiration aesthetics (Abidin, 2016; Hayes & Ben-Shmuel, 2024). Marketing and communication research formalized the mechanisms behind influence. Sokolova and Kefi (2020) and Lou and Yuan (2019) showed how credibility, authenticity, and source congruence drive persuasion. Reinikainen et al. (2020) and Tafesse and Wood (2021) provided evidence of the greater engagement and adherence to suggestions with the help of parasocial bonds and community rituals. Therefore, accounts-level analyses and event studies, embraced in measuring attention spillover, returns patterns, and welfare of investors, have been undertaken in both the fields of finance and accounting fields, with cryptocurrency and meme stocks serving as natural laboratories (Cookson et al., 2023; Barber et al., 2022; Ante, 2023; Merkley et al., 2024; Warkulat & Pelster, 2024). Another strand of thought is to investigate reactions of policies and derive the guiding principles that can be relevant to all important markets, despite the inconsistency in procedural requirements and approvals (ESMA, 2024).

Two integrative insights have occurred. First, authenticity is strategic. It is learned and performed because platforms reward content like sincere peer conversation (Mölders et al., 2025; Sokolova & Kefi, 2020; Zhu et al., 2019). Second, platform design is causal. Ranking, novelty, and watch time incentives shape which financial messages travel far and how fast frames diffuse (Shin, 2022; Vosoughi et al., 2018; Appel et al., 2020). These insights imply that policy should address not only bad actors but also design choices by firms and platforms that govern how disclosures and risks are presented (Baek et al., 2010).

Access, Appeal, and Accountability

Findings across fields can be organized into a three-factor model. Access describes the collapse of distribution costs, allowing one video to reach millions. Appeal captures the persuasion stack of story, strategic authenticity, and parasocial bonding that converts attention into action. Accountability describes the evolving rules that carry classic truthfulness, balance, and conflict management principles into social feeds and allocate responsibility across creators and firms (Krämer et al., 2015). Outcomes depend on how these forces interact. When access and appeal operate without accountability, speculative promotions flourish and investor welfare deteriorates. When accountability reins in conflicts and creators commit to education-first narratives, access and appeal widen entry into basic planning and raise the floor of literacy at scale (Barber et al., 2022).

Unresolved Questions and Research Gaps

Four gaps stand out. First, long-run household outcomes remain undermeasured relative to short-run market reactions. Linking privacy safe exposure data to account-level outcomes would enable stronger causal claims (Barber et al., 2022; Warkulat & Pelster, 2024). Second, algorithmic causality still relies on inference. The cooperation with platforms to pilot ranking and label designs would work out how messages propagate (Shin, 2022; Vosoughi et al., 2018; Cotter, 2019). Third, cross-border variances in disclosure, approval, and licensing are a factor that prompts systematic mapping to establish regulatory mixes that reduce harm without suppressing valuable education (ESMA, 2024). Fourth, creator economics will likely bias the selection of topics and assertion strength. The comprehension of sponsorship and affiliate incentives would make the ethical

guidance more tangible and allow supervisors to focus on the riskiest areas of the funnel (Cornwell & Kwon, 2020).

The literature portrays finfluencers as translators making money talk feel like part of ordinary life and promoters whose reach and speed can amplify speculative behavior. Education First channels align with the diversified, low-cost strategies that support household welfare. Trading first scripts pushes followers toward higher risk and higher cost behaviors that reduce subsequent returns. Regulators are translating long-standing principles into guidance for feeds, and jurisdictions are converging on disclosure, balance, and responsibility expectations. The next steps for research are empirical and collaborative. Long horizon measures of household outcomes and transparent tests of platform design choices are necessary.

Research Design and Rationale

In this section a desk research design is used to synthesize what is known about financial influencers and the circulation of investment and personal finance advice on social platforms. Desk research is appropriate when the object of study moves quickly across jurisdictions and platforms and where primary fieldwork would be outpaced by fast publication cycles and policy changes. An integrative review logic was applied to synthesize a coherent account from diverse empirical and conceptual studies, while a scoping review frame was embedded to map the breadth of evidence, definitions, and measures within this relatively young field. The aim is twofold. First, to describe who finfluencers are, what they do, and how they communicate. Second, to evaluate how regulators and professional bodies have responded and what the academic evidence suggests about potential benefits and harms for consumers.

An integrative desk review allows the inclusion of qualitative studies, survey-based evidence, experiments, and computational analyses that examine related phenomena such as attention dynamics, parasocial relationships, and the spread of misinformation in financial contexts. At the same time, a scoping approach is suitable for heterogeneous literatures where basic mapping of concepts, methods, and gaps is required before any formal meta-analysis is feasible. Because policy and supervisory guidance strongly shape the finfluencer ecosystem, the desk research also includes documentary analysis of publications from major market regulators and standard setters. These sources are treated as reputable non-academic evidence and appraised transparently to avoid over-weighting advocacy or communications material.

Protocol Registration and Reporting Standards

A methods blueprint was drafted a priori that specified the research questions, eligibility criteria, search strategy, screening workflow, quality appraisal tools, and plan for dealing with possible bias. The reporting follows PRISMA 2020 guidance for transparent synthesis of evidence, with the search process documented according to PRISMA S for search reporting (ESMA, 2024). Although scoping reviews are not always registered on clinical registries, documenting the methods and noting any deviations that arose during screening and extraction follows the spirit of protocol registration. When synthesis proceeds without pooling effect sizes, the SwiM recommendations are applied to describe how decisions were made regarding grouping studies and summarizing directions of effect (ESMA, 2024).

Eligibility Criteria

The studies were considered eligible as long as they covered any of the following areas: (i) how influencers or content creators on YouTube, TikTok, Instagram, X, Reddit, or podcasts create, seek, or receive investment advice or personal wealth advice; (ii) the behavior of retail investors, who follow asset-related or personal financial content via social media; (iii) the strategic plans, authenticity practices, or disclosure strategies to which influencers rely on when talking about financial matters; or (iv) regulatory frameworks, laid out by financial supervisory bodies, to regulate their activities. Eligible academic designs included qualitative interviews or ethnography, content analysis, experiments, surveys, digital trace or platform data studies, and systematic or scoping reviews focused on finance advice or closely adjacent domains such as crypto promotion and stock market commentary. Exclusions were set for articles that studied influencer marketing unrelated to finance, platform policy commentary without an analytic method, opinion pieces without data, and studies focused exclusively on firm-led financial marketing unless they contained findings directly transferable to independent creators.

Temporal and linguistic boundaries were introduced to keep the corpus contemporary and manageable. The principal window ran from January 2015 to July 2025. English language items were included due to resource constraints in translation, while acknowledging that significant communities operate in other languages. In the non-academic strand, eligible documents were those that were issued by a recognized supervisory authority or professional standard setter with jurisdictional authority over financial promotions, including IOSCO, the FCA, ESMA, SEC, ASIC, SEBI, BaFin, and the Monetary Authority of Singapore. These sources have been included since regulatory documents define compliance requirements, clarify the enforcement priorities, and

provide market risk assessments, which inform influencer behavior in the market and the policies governing platforms.

Sources and Search Strategy

The academic search strategy combined subject database searches with citation chasing. Computational studies measuring content diffusion or algorithmic exposure were found in the core databases: Scopus, Web of Science Core Collection, EBSCO Business Source, PsycINFO, and IEEE Xplore. These resources are commonly suggested to be searched under social science and business topics.

For regulatory and policy documents, targeted site limited searches were run on regulator domains and the IOSCO, ESMA, and FCA publications portals. Although these are not peer-reviewed journals, they constitute reputable sources under the documentary analysis strand and were handled with a separate appraisal tool, described below.

Screening and Study Selection

Disagreements were resolved by revisiting the protocol and re-reading abstracts. Full text screening followed, and reasons for exclusion were applied according to the PRISMA flow diagram. As suggested by PRISMA 2020, study selection was piloted on a small batch to calibrate decisions before screening the full set. The same logic was applied to regulator documents, with additional attention to document type because news releases, blogs, and speeches do not always have the same normative status as rules or formal guidance.

Quality Appraisal

Given the diversity of designs in this field, no single appraisal tool suffices. The approach was tailored to the study type and aligned with precedent in mixed-evidence reviews. For empirical mixed evidence, the Mixed Methods Appraisal Tool version 2018 was used to assess core quality features consistently across qualitative, quantitative, and mixed methods studies. For randomized and quasi-experimental designs, risk of bias was considered along domains analogous to RoB 2, with attention to deviations from intended interventions, confounding, measurement of outcomes, and selective reporting. For qualitative work, trustworthiness was assessed using credibility, transferability, dependability, and confirmability criteria. For policy and regulatory documents, the AACODS checklist was applied to evaluate authority, accuracy, coverage, objectivity, date, and significance in the absence of peer review.

Quality ratings were not used to exclude all lower-scoring studies because this would erase important signals in an emerging literature. Instead, ratings informed the weight placed on findings in the synthesis and the confidence statements offered in the discussion.

Addressing Bias, Reliability, and Reflexivity

Desk research carries familiar validity concerns. Publication bias may privilege significant or novel findings. Platform access limits and changing application programming interfaces can bias digital trace studies toward observable communities. To mitigate these problems, the search strategy was intentionally broad across disciplines, citation chasing was used to surface adjacent work, and grey but reputable regulatory documents were included with transparent appraisal to reduce the risk that policy evidence is considered only through secondary commentary. The PRISMA S documentation

of search terms and database coverage supports reproducibility and helps readers judge whether material gaps could materially change the conclusions. The study's orientation is directed toward investor protection and information quality. Accordingly, the method examines beneficial and harmful effects and evaluates communication practices against established advertising and disclosure research rather than relying solely on platform norms.

Ethics and Research Governance

This project did not involve direct interaction with human participants or collecting personal data. It is therefore outside the scope of institutional human subjects review in many settings. Nevertheless, basic ethical considerations apply to the handling of sensitive topics and the responsible representation of content creators. Only information available in published studies and official regulatory documents was used. Care was taken to avoid reproducing defamatory claims or identifying individual creators in ways that are not necessary for analysis. The synthesis favors generalizable patterns over commentary on specific persons or accounts.

Methodological Limitations

Several limitations arise from the chosen design. First, limiting sources to English-language publications may underrepresent important communities in markets where influencer activity is vibrant. Second, reliance on publicly available regulator documents can miss supervisory practices that are not published. Third, desk research cannot confirm causality where observational studies correlate exposure and behavior. Finally, platform and policy landscapes change quickly. Even with late-stage updates, findings may age faster than in slower-moving fields. To address these constraints, the synthesis emphasizes mechanisms and communication patterns that are likely to

persist across minor platform changes, and it links those patterns to regulatory principles that are themselves relatively stable.

Results and Discussion

The desk review provided a multi-strand body of work that cut across communication, marketing, information systems, behavior, financial, and cultural economy. A similar opinion arose among the various fields that financial influencers belong to an intermediate space between media personalities and informal educators who act as market promoters (Symbiosis & Gandhi, 2024; Haase et al., 2023). This hybridity materialized in how they performed their roles in the area, the contents and formats they adopted, and the reactions of audiences and regulators. The results are presented in thematic sections and discuss the implications for theory/practice and policy.

What Finfluencers Are and How They Are Positioned

Recent research conceptualized financial influencers as creators who generate audiences based on investment and personal finance content and market authenticity and expertise in manners configured to platform culture (Abidin, 2016; Hayes & Ben-Shmuel, 2024; Ki et al., 2020). With ethnographic and cultural economic accounts, the persona was not just a medium to transfer the information. However, it can be seen as a format to fashioned identity that combines lived experience, aspiration narratives, and proficiency claims that frequently incorporate a focused presentation as a learning process or on service to the community (Abidin, 2016; Hayes & Ben-Shmuel, 2024; Djafarova & Bowes, 2021). Management and marketing also defined influencers as human brands capable of synthesizing complicated financial concepts into scripts and repeatable heuristics that help to bridge the gap between professional advice and word of mouth among peers (Ki et al., 2020; Casaló Alberto et al., 2020; Hudson et al., 2015).

Typologies included educator-setup mentors, product-reviewers or affiliates, market chatters, and trader-entertainers focusing on action and excitement (Sousa et al., 2025). Strategic authenticity emerged as a fundamental organizing logic in that creators show transparency in losses and gains, portray behind-the-scenes work, or feature everyday life to foster intimacy and trust-creation with the authenticity of user-generated artists, as companies, could renege on commitments to maximize profits, following the previous logic of being authentic to reduce and avoid conflicts. These identity and role relations place influencers at the intersection of the formation of parasocial relationships and perceived expertise, the two modalities that drive the way the audiences perceive and transform the messages relayed by the former (Jin et al., 2019; Hwang & Zhang, 2018; Reinikainen et al., 2020).

Communication Formats, Strategies, and Engagement Devices

Across platforms, financial influencers relied on a repertoire of communication strategies that research has associated with influencer effectiveness more broadly (Johnstone & Lindh, 2018; Leung et al., 2022). Recurrent features included narrative framing, simple visual analogies, personal testimony, and call-to-action prompts that lower the perceived threshold to act, such as "start with ten dollars" or "open a demo account" (Campbell et al., 2020; Lou et al., 2019). Content analysis and audience studies showed that disclosure of sponsorships, when present, often appeared in subtle forms or in locations that reduced salience, with limited activation of persuasion knowledge among less experienced viewers (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; van Dam & van Reijmersdal, 2019; Cornwell & Kwon, 2020).

Parasocial cues and community practices were central to engagement. Findings revealed the importance of conversational tone, regularity of posting, mutual interactions in the comments, and references to a shared journey in reinforcing a sense of closeness, which then predicts trust and adherence to suggestions (Jin et al., 2019; Reinikainen et al., 2020; Tafesse & Wood, 2021). Multimodal postings with expressive imagery combined with succinct text overlays and elements unique to specific platforms, such as stitches or duets, were more likely to receive reactions and shares that increased reach by being favorited by recommendation algorithms (Rietveld et al., 2020; Appel et al., 2020; Thorson et al., 2021). The use of short-format videos led to greater visibility and repeat exposure, in line with what has been found about mobile feeds and the dopaminergic involvement of social media consumption (Montag & Hegelich, 2020; Cinelli et al., 2020).

A further strand of evidence tracked how authenticity cues and credibility signals interact. Claims of independence, experience-based teaching, and selective admission of mistakes-built authenticity, whereas the presence of affiliate links, broker partnerships, or crypto token incentives raised perceived conflicts of interest unless transparent and justified (Belanche et al., 2021; Yu & Lee, 2019; Sousa et al., 2025; Djafarova & Rushworth, 2017). The credibility literature suggested that fit between creator persona and topic increases persuasion, with micro influencers sometimes outperforming macro figures on perceived trust due to niche expertise and stronger community bonds (Kay et al., 2020; Ki et al., 2020; Djafarova & Rushworth, 2017).

Topics and Product Categories

There are three major areas in which the thematic focus of influencer content gathers. The first area concerns the basics of personal finance-budgeting, saving, credit building, and long-term investment using index funds or diversified portfolios (Khurana, 2023; Hammer, 2025). They often take a peer-like position in this arena, offering easy entry points and de-jargoning technical terminology. They often refer to the traps of behavior and simplistic rules of thumb (Abidin, 2016). The second domain involved trading and asset picking, most intensively around equities and options during periods of elevated retail participation, with creators demonstrating strategies, highlighting news catalysts, or reacting to earnings (Cookson et al., 2023; Barber et al., 2022; Chadwick et al., 2018). The third domain concerned crypto assets and Web3, where creators spanned education, project promotion, and market commentary. Work on crypto influencers showed systematic patterns in how token promotion and community hype relate to attention and trading, with risks heightened by opacity and the speed of cycles (Ante, 2023; Merkley et al., 2024; Kedvarin & Saengchote, 2023).

Attention dynamics recurred across these topics. Studies of social media attention and retail trading suggested that spikes in creator coverage and platform engagement often preceded or coincided with flows from retail investors and short-horizon price drift, a phenomenon consistent with attention-induced trading (Warkulat & Pelster, 2024; Barber et al., 2022; Krämer et al., 2015). The rationality behind it is salience and simplified messages that decrease cognitive load, thus creating action in terms of less deliberative users (Pennycook et al., 2021). Although some content makers have supported long-term, diversified approaches, increased engagement always accumulates to content that makes possible novelty, urgency, and a feeling of community belonging, thus dictating processes of online diffusion (Vosoughi et al., 2018).

Observed and Inferred Outcomes for Audiences

Empirical data on results have shown great heterogeneity in creator typologies, audience levels of literacy, and product risk exposures. On the bright side, intimidation can be alleviated, the culture of basic saving can be normalized, and investment-literacy principles can be taught to young adults at earlier stages than have historically been made available through the traditional media (OECD, 2021; Abidin, 2016; Appel et al., 2020). Engagement is maintained via peer-to-peer social learning and perceived relatedness beyond single financial campaigns and parasocial relationships enhance the motivating effects (Jin et al., 2019; Hwang & Zhang, 2018). When the messages of the creators are aligned with the consumer-finance best practices in the low-risk situations, like budgeting and emergency funds, the convergence of the two can imply the possibility of scalable returns (Malyavkina, 2018; OECD, 2021; Hoffman & Novak, 2018).

Risks accumulated as content shifted toward complex or speculative products. Quantitative finance research reported that retail attention can lead to concentrated flow in popular tickers with transitory return patterns and heightened intraday volatility, consistent with noise trading and limits to arbitrage in the face of attention shocks (Barber et al., 2022; Cookson et al., 2023). In crypto markets, creator announcements and endorsements were correlated with abnormal social media activity and short window price effects, while subsequent underperformance or fraud events-imposed losses on late-arriving followers (Ante, 2023; Merkley et al., 2024; Erkan & Evans, 2018; Krause, 2025). Information quality concerns persisted. Work on misinformation spread and on the limits of disclosure recognition indicated that many users fail to detect sponsorship cues or to verify claims when attention and emotion are high (Boerman & Van Reijmersdal, 2020; Vosoughi et al., 2018; Pennycook & Rand, 2019; Shan et al., 2020).

Credibility, fit, and perceived authenticity mediated audience trust and behavioral compliance (Yılmazdoğan et al., 2021). Experiments and surveys indicated that when creators were perceived as an expert and genuine, the desire to act on the information was higher, especially when dealing with content that had a relationally framed message (Ki et al., 2020; Reinikainen et al., 2020). Macro may have a lower engagement rate than micro influencers, but t, they provided a scale that magnified the positive or negative effect (Kay et al., 2020; Tafesse & Wood, 2021).

Ethical and Regulatory Responsibility

The ethical landscape took shape around the issue of conflict of interest, suitability, and vulnerability. Research on influencer marketing consistently indicated that much of the disclosures can be skimmed over or misunderstood, and that advisory messages can easily be confused with

persuasive ones when a creator explicitly expresses a call to take action on a particular product, open an account, follow a strategy (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; van Reijmersdal et al., 2020; Vrontis et al., 2021). In the financial context, this blurring has greater implications due to the non-triviality of losses that can be suffered by investors, as well as because such investors include minors and low literacy populations in the captive audiences of highly entertaining material (De Jans et al., 2018; Hudders et al., 2021).

The concern was manifested in the regulatory texts of the past few years. Supervisors emphasized that financial promotions on social media must be fair, transparent, and non-misleading and that risk alerts, advantages, and risks must be fairly displayed with equal importance (ESMA, 2024; Kizgin et al., 2018). Several regulators have also reminded the regulated companies that they have obligations under combined efforts with creators, including due diligence, record keeping, and responsibility in promoting high-risk products (Mason & Clarke, 2025). The sector's regulators also warned against the participation of licensed entities in establishing unregistered influencers to avoid regulatory arbitrage and pseudo-independence (Singh & Sarva, 2024). Trans-boundary comparisons had revealed common features: disclosure of incentives, prominence of risk warnings, and not using undue pressure, although the line between education and promotion would always still need to be considered on a case-by-case basis (Mason & Clarke, 2025).

Efficacy as Financial Advisors and Performance Considerations

Whether finfluencers function effectively as advisors depends on the benchmark. Most creators work outside of formal assessment processes when compared against professional suitability standards and fiduciary obligations (Lalwani, 2025; Ben-Shmuel et al., 2024; Burgess, 2025).

Some were cautious and diversified, but many focused-on strategies with entertainment value or short-horizon payoff framing, which can lead to poor timing and excessive followers' turnover (Barber et al., 2022; Cookson et al., 2023). The attention literature indicated that price reactions around high-visibility content are often short-lived and can reverse, implying that chasing creator-driven signals may not produce persistent alpha after costs (Warkulat & Pelster, 2024).

However, efficacy is not confined to stock picking or trade timing. Communication research demonstrated that creators can be efficient at taking initial actions of long-term value, like opening a retirement account, creating an emergency fund, or automating savings, by reducing psychological barriers, providing social proof (Lou & Yuan, 2019; Ki et al., 2020; Appel et al., 2020). Human branding studies and authenticity literature claimed that by being exposed to voices of trust over multiple applications, prudent practices can be normalized and habit formation can go on a scaffolding effect of being bombarded by trusted voices, even when specific market calls are not superior in their performance in contrast to chance (Campbell et al., 2020; Yu & Lee, 2019; Hudson et al., 2015). The implication is that performance should be evaluated along two axes: market timing accuracy and behavioral change. Finfluencers perform unevenly on the first axis and more consistently on the second when their content aligns with basic financial literacy principles (Johnstone & Lindh, 2018; Krause, 2025).

Platform Dynamics and Algorithmic Mediation

Another layer of results concerned platform architecture. Algorithmic recommendation and engagement-based ranking tended to privilege content with high watch time, strong early interactions, or salient novelty, which often favor bold claims and simple narratives (Appel et al.,

2020; Thorson et al., 2021). Short video ecosystems created strong path dependence in exposure; once a user engaged with finance content, the feed supplied more of the same, intensifying the salience of that domain (Yen et al., 2024; Khandolkar et al., 2024). In such environments, disclosure prominence and risk framing compete with attention optimized devices, so subtle warnings or nuanced caveats lose out to kinetic editing and declarative statements (Boerman & Van Reijmersdal, 2020; Shan et al., 2020). These platform-mediated forces complicate individual responsibility framings and support policy approaches that address format-level risks.

Cross-Cutting Discussion and Implications.

Taken together, the results point to three cross-cutting insights. First, finfluencers thrive because they solve a real communication problem. Financial services communication has long struggled with jargon, low perceived relevance, and a trust deficit. Creators reverse those disadvantages through persona-driven pedagogy, community reciprocity, and narrative devices that make saving and investing possible and enjoyable (Ki et al., 2020; Reinikainen et al., 2020; Campbell et al., 2020). The cultural economy lens shows that this is not an accident; it is the product of labor at the intersection of self-branding and vernacular expertise (Abidin, 2016; Hayes & Ben-Shmuel, 2024; Zhu et al., 2019).

Second, the same features that make finfluencers engaging also generate risk when incentives and product risk escalate. Strategic authenticity and parasocial proximity can take audiences beyond the educational premise and straight into doing without proper deliberation and due checks about suitability (Jin et al., 2019; Hwang & Zhang, 2018; Lalwani, 2025). The dynamics of attention could shift the prices at the margin and crowd the retail flow, with the potential to make late

followers suffer losses and result in self-reinforcing movements in illiquid assets (Barber et al., 2022; Cookson et al., 2023). The weak or vague disclosures add to the problem, more so in feeds that are rewarded by the speed and brevity of the disclosures (Boerman & Van Reijmersdal, 2020; van Dam & van Reijmersdal, 2019).

Third, the regulatory response converges on principles that, if operationalized, can preserve benefits while reducing harm. There are common priorities of clear and prominent risk warnings, plain language risk and reward balance, transparent disclosure of incentives and affiliate relations, and strong oversight of firm creator partnerships (ESMA, 2024). The advertising research evidence can be provided with concrete advice to implement: disclosures should be timely, prominent, and unambiguous; risk statements should be co-located with the claim; and the format should not be designed in a way that compromises understanding (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; Kim & Kim, 2021). Platform design remains a crucial partner lever. Where platforms enforce financial services and policies and elevate disclosure prominence, user comprehension improves (Thorson et al., 2021; Appel et al., 2020).

These implications also make clear that some forms of communication should be limited. For example, explicit and specific investment advice, that would already be generally not preferable, should always be accompanied by clear and equally explicit risk disclosures that assume the same prominence as the main message. Seeking this balance between content of the message and implied risks is justified by the fact that if concentrated risks and information asymmetries are structural characteristics of the users' structure then solutions should be structural as well.

Limitations of The Evidence and Research Agenda

The literature has matured rapidly, but several gaps remain. First, causal inference about advice effectiveness is difficult outside randomized designs, and natural experiments using platform policy changes or sudden creator suspensions remain scarce. Studies that combine field experiments with platform-level collaboration would strengthen estimates of behavioral impact. Second, more work is needed on non-English communities, given the scale of influencer activity in Asia, Latin America, and non-English Europe. Third, long-horizon outcomes, such as persistence of saving habits or portfolio quality improvements, are understudied relative to short-horizon trading effects. Fourth, disclosure science in finance-specific contexts would benefit from large-scale A/B tests that vary placement, wording, and visual salience across formats, informed by what is known from influencer marketing but tuned to financial risk (Boerman & Van Reijmersdal, 2020; van Reijmersdal et al., 2020; Mason & Clarke, 2025; Masuda et al., 2022). Finally, the presence of creator networks, copy-trading prompts, and social trading websites outlines a frontier where individual and collective behavior overlap; computational social science could be used to provide a more accurate mapping of these networks and the resultant price impact (Cookson et al., 2023; Zhang & Gong, 2021; Martínez-López et al., 2020).

Practical and Policy Recommendations

Several practical measures have been proven by evidence. With creators who pursue education opportunities and reduce threats to any group, the focus on low-cost diversified strategies and the application of the basic principles of literacy can contribute to the delivery of value to the majority of the population and minimize harm (OECD, 2021; Malyavkina, 2018). In companies that utilize creators, due diligence exercises on disclosure practices, reviewing the content to ensure balance,

and training providers on the regulators' expectations would mitigate the risk of enforcement actions (Singh, 2025). One recommended way on platforms is to standardize online disclosure tools, making them salient in viewing and imposed on financial-related information (Thorson et al., 2021; Appel et al., 2020). Providing guidance to supervisors (with specific examples of compliant and non-compliant posts) and more guidance on the distinction between education and promotion would enhance certainty and compliance (ESMA, 2024).

The results Indicated that the finfluencer phenomenon emerged at the Intersection of unmet communication needs in consumer finance, platform architectures that reward engaging narratives, and a regulatory perimeter that initially lagged behind practice. Finfluencers can democratize access to financial concepts and motivate constructive action, especially for early-stage behaviors. They can also expose audiences to conflicts of interest, speculative strategies, and misleading framing that erode consumer protection goals. The discussion argued that the task is not to suppress creators but to shape incentives and guardrails so that the persuasive power of persona-driven finance improves household outcomes rather than undermines them. Such coordination will need creators, firms, platforms, and regulators to work together based on the emerging body of literature on topics of authenticity, disclosure, attention, and investor behavior (Campbell et al., 2020; Boerman, 2020; Barber et al., 2022; Cookson et al., 2023).

Recommendations

The evidence suggests that financial influencer activity delivers public value and meaningful risk. Recommendations, therefore, emphasize a twin goal: preserve the motivational and pedagogical strengths of creator-led finance while reducing the probability and severity of consumer harm. The

guiding principles include transparency, proportionality, and alignment between persuasive formats and the comprehension needs of retail audiences (Hull & Qi, 2024). These principles are aligned with the advertising disclosure literature, which stresses the importance of prominence, timing, and clarity, as well as with investor protection regulations, which require communications to be fair, clear, and not misleading (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; van Dam & van Reijmersdal, 2019; ESMA, 2024). More in general, what emerges is that principles of autonomy, nonmaleficence, and fairness should be operationalized directly in the platforms and integrated in the feed-based environment that is specific to social media.

Recommendations for Regulators and Standard Setters.

Clarify the boundary between education and promotion

Supervisory guidance should adopt a functional test that focuses on call to action, specificity, and the presence of incentives. A post that names a product, urges acquisition or account opening, or embeds affiliate remuneration should be treated as a promotion regardless of educational framing. A post that explains concepts without directing a transaction can be treated as education, with safe harbor conditions that include the absence of compensation and the use of balanced examples (ESMA, 2024). Disclosure science shows that audiences miss or misinterpret subtle cues, which argues for clear categorical thresholds rather than intent tests alone (Boerman & Van Reijmersdal, 2020; Kim & Kim, 2021).

Standardize risk presentation and disclosure form factors

Rules should require that risk statements appear with the same salience, spatial proximity, and audiovisual weight as benefit claims, including in short video and story formats. Research on

disclosure effectiveness supports simple language, early placement, and repetition across frames where content is segmented (Boerman & Van Reijmersdal, 2020; Shan et al., 2020). Creators who receive value should disclose the nature of compensation concisely, while regulated firms that sponsor content should be responsible for the accuracy and balance of the message (Wellman et al., 2020).

Set obligations for partnerships between firms and creators

When licensed entities work with creators, obligations should include due diligence on the creator's compliance history, preapproval and archiving of content, ongoing monitoring, complaint and redress pathways, and recordkeeping sufficient to evidence supervisory compliance. These obligations mirror existing financial promotion requirements but adapt them to social formats and creator workflows (Singh, 2025).

Target high-risk product promotion with proportionate restrictions

Promotions for complex or speculative instruments such as high-leverage derivatives and unbacked crypto assets should be constrained to contexts where audience suitability checks are feasible (Kedvarin & Saengchote, 2023). Where such promotions are allowed, enhanced warnings, cooling-off periods, and risk acknowledgement prompts are recommended, backed by attention research showing action spikes under novelty and urgency (Barber et al., 2022; Cookson et al., 2023; Ante, 2023; Merkley et al., 2024).

Encourage data access and transparency

Regulators should promote or require public ad libraries for finance promotions, including creator content sponsored by firms, with searchable metadata on sponsor, spend, impressions, and targeting where applicable. Such libraries support monitoring and research and align with calls for transparency in algorithmic curation and advertising (Thorson et al., 2021; Appel et al., 2020; De Veirman & Hudders, 2020). Memoranda of understanding for cross-border cooperation to align responses to transnational promotion campaigns (Thorson et al., 2021).

Adopt evidence-informed enforcement

Enforcement is best targeted at egregious deception, undisclosed conflicts of interest, and persistent non-compliance, whilst providing corrective routes for low-harm cases where guidance and training can quickly improve practice (SEC, 2022). Insights from the research indicate that many failures in comprehension are a matter of format and position rather than of conscious deception, which supports a tiered approach (Boerman & Van Reijmersdal, 2020; van Dam & van Reijmersdal, 2019).

Recommendations for Platforms

Build finance-specific disclosure and risk modules into native tools

Platforms should provide standardized overlays for paid partnership and affiliate relationships that remain visible across all viewing contexts and are resilient to cropping or remixing. Prompts for risk statements should be integrated into upload flows when finance is detected, with structured fields that enforce plain language and sufficient length. Such design choices are supported by

evidence that disclosure prominence and timing shape recognition (Boerman, 2017; Kim & Kim, 2021; Shan et al., 2020).

Strengthen identity and eligibility checks for finance promotions

Paid promotion tools that can reach large audiences should be available only to verified advertisers, subject to policy review. Identity checks and documentation of regulatory status or firm sponsorship reduce room for astroturfing and misrepresentation. This aligns with regulatory expectations and platform governance work that ties accountability to verified identity in sensitive domains (ESMA, 2024; Thorson et al., 2021).

Introducing accuracy and reflection nudges in finance contexts

Accuracy prompts that ask creators to consider whether a claim is supported and invite viewers to reflect before acting on a recommendation have been shown to improve attention to truth without heavy-handed friction (Pennycook & Rand, 2019; Pennycook et al., 2021). In finance, such prompts can be triggered when claims include specific tickers, tokens, or guaranteed returns.

Address algorithmic amplification of high-risk cues.

Engagement-based ranking prefers dramatic, novel, and urgent content, which may elevate speculative narratives (Appel et al., 2020; Thorson et al., 2021). Platforms should test down-ranking of finance content that contains promissory language, hidden or minimal disclosures, or repeated violations, alongside the elevation of educational content that meets disclosure and balance standards. The misinformation literature provides evidence that design decisions can mitigate the frictionless spread of misleading information while preserving legitimate discourse (Vosoughi et al., 2018; Roozenbeek & Van der Linden, 2019).

Support researcher access and public accountability

Privacy-preserving data access for accredited researchers studying finance content, disclosure effectiveness, and audience outcomes would enable more rigorous independent evaluation. The goal is to measure and reduce externalities such as herding and attention-induced trading shocks in a manner consistent with consumer protection (Barber et al., 2022; Cookson et al., 2023; Warkulat & Pelster, 2024).

Recommendations for Financial Firms and Industry Bodies

Adopt partnership governance frameworks

Firms that engage creators should institutionalize end-to-end processes: vetting, training on applicable rules, preapproval of scripts or key messages, version-controlled asset libraries for risk statements, content archiving, and post-publication monitoring. Contracts should require the conspicuous disclosure of the relationship and bar claims that evidence cannot substantiate. These measures turn regulatory expectations into operational controls (Singh, 2025).

Align campaigns with behavioral evidence

When the educational goal is to foster saving and diversification, campaigns should use narrative and peer modeling approaches shown to enhance engagement and trust while discouraging the glamorization of short-horizon trading (Ki et al., 2020; Lou & Yuan, 2019; Appel et al., 2020). Where product promotion is appropriate, firms should ensure that balancing information is not buried and that suitability pathways are clearly signposted (ESMA, 2024).

Measure comprehension and not only reach

Campaign evaluation should include metrics that capture recognition of disclosures, recall of risks, and comprehension of key facts, not only impressions and clicks. The advertising literature shows that comprehension mediates downstream persuasion effects and should, therefore, be a primary performance indicator (Boerman & Van Reijmersdal, 2020; van Reijmersdal et al., 2020). Pre-testing creative variants can identify effective balance and phrasing before wide release.

Engage in industry codes and third-party oversight

Trade associations can maintain voluntary standards for creator partnerships above the legal minima, including model disclosure language, negative lists of prohibited claims, and mutual audit protocols. Independent review or certification can increase baseline quality and enable public trust (Hudders et al., 2021; De Jans et al., 2018).

Recommendations for Creators

Adopt an audience-first disclosure ethic

Disclosures should be brief, plain, and repeated across frames where content is segmented. The form should state the sponsor and the nature of compensation, for example, “paid partnership with X” or “affiliate links generate commission,” placed early and spoken in audio where possible. Evidence has indicated that early and salient disclosure can improve recognition without reducing engagement when the content is valuable (Boerman & Van Reijmersdal, 2020; Evans et al., 2017; Kim & Kim, 2021).

Balance benefits and risks with concrete examples

When discussing products or strategies, creators should accompany claims of possible benefit with illustrative risks and realistic ranges of outcomes (Miettinen, 2025). Balanced presentation helps understand and mitigate the perception of guaranteed returns, an important issue for investor protection (ESMA, 2024). Educational content can be grounded on principles such as diversification and cost control that are well supported in the literature (OECD, 2021).

Use authenticity responsibly

Strategic authenticity is a trust builder, but it should not be used to downplay risk or suggest expertise beyond competence. Sharing learning experiences and mistakes is valuable. It can be combined with clear boundaries, for example, not telling them what to trade and encouraging them to conduct their own research or consult licensed professionals where relevant (Jin et al., 2019; Reinikainen et al., 2020; Hayes & Ben-Shmuel, 2024; Zhu et al., 2019).

Avoid manipulative attention devices

Promissory or urgency-laden titles and thumbnails, backtested performance presented as certain, and omission of fees or taxes should be avoided. Such devices may enhance short-term engagement but undermine trust, and heighten the risk of harm, as well as being associated with short-term spikes in attention that reverse (Barber et al., 2022; Cookson et al., 2023; Warkulat & Pelster, 2024).

Correct errors promptly and maintain an archive

Errors to be recognized and corrected in the channels in which the original content was published. Publicly archiving corrections and sponsored posts makes them accountable and aligns with

audience expectations in education-focused communities (Hudders et al., 2021; Belanche et al., 2021; De Veirman & Hudders, 2020).

Recommendations for Educators, Civil Society, and Public Agencies

Partner with credible creators to scale financial literacy

Programs can create modular content that describes budgeting, emergency funds, diversified investing, and fraud avoidance. Narrative and peer-based formats are more engaging and perceived as relevant than traditional didactic formats (Ki et al., 2020; Appel et al., 2020; Tafesse & Wood, 2021). Co-branding should not undermine autonomy or provide or even give the impression of endorsing any particular products (Arora et al., 2023).

Teach recognition of disclosures and risk cues

Public literacy campaigns should include basic exercises that make people more aware of sponsored content and standard persuasion techniques. Finally, research has found that providing users with cues and prompting short-term accuracy reflection can mitigate false belief vulnerability (Boerman & Van Reijmersdal, 2020; Pennycook & Rand, 2019; Pennycook et al., 2021; Wellman et al., 2020).

Develop rapid response guidance for emerging narratives

During market manias or novelty cycles, authoritative guidance can serve as a calm, balanced context to which creators and journalists can refer. The misinformation literature confirms that it is better to debunk misinformation promptly and with accurate alternatives than with blanket warnings (Vosoughi et al., 2018).

Recommendations for Researchers

Prioritize causal identification and long-horizon outcomes

Natural experiments that take advantage of platform policy changes, staggered introductions of disclosure tools, or creator suspensions can help tease apart effects on behavior and price formation. Longitudinal studies should monitor continuous changes in saving behaviour and portfolio quality, not just short-window trading (Barber et al., 2022; Cookson et al., 2023; Warkulat & Pelster, 2024).

Advance disclosure science in finance contexts

Large-scale A and B tests can evaluate placement, wording, and multimodal delivery of finance-specific disclosures, including risk statements, affiliate explanations, and performance disclaimers. The influencer and advertising literatures provide starting points, but finance introduces different stakes and comprehension needs (Boerman & Van Reijmersdal, 2020; Shan et al., 2020).

Broaden linguistic and cultural coverage

Non-English communities are underrepresented in work to date. Cross-linguistic and cross-regulatory comparative research would help to shed light on how norms and regulations influence creator practice and audience reception (Hudders et al., 2021; De Jans et al., 2018; Hayes & Ben-Shmuel, 2024).

Map creator networks and amplification

Computational studies should trace network structures, content flows, and co-movement between attention and trading to distinguish organic education communities from coordinated promotion. Prior work on attention-induced trading and diffusion can be extended with richer network measures (Zhang & Gong, 2021; Rietveld et al., 2020).

Pursue open science and research access

Preregistration, shared codebooks, and data will enhance comparability and cumulation. Working with platforms and regulators enables privacy-respecting access to data required for robust inference (Thorson et al., 2021; Appel et al., 2020).

Monitoring, Metrics, and Evaluation

Focus on comprehension and behavior, not only exposure

Monitoring should include measures of disclosure recognition, risk recall, comprehension of core facts, and balanced perception gathered through experiments or surveys embedded in platforms (Young et al., 2018). These variables mediate persuasion and predict action (Boerman & Van Reijmersdal, 2020; Kim & Kim, 2021). Behavioral metrics should include adoption of basic financial practices, not only short-horizon trading (Singh, 2025).

Track market externalities

At the market level, attention-driven herding indicators, abnormal turnover in retail heavy names, and short window reversal can inform supervisory dashboards. Prior finance research offers

methods to link social attention and flows (Barber et al., 2022; Cookson et al., 2023; Warkulat & Pelster, 2024).

Evaluate policy interventions iteratively

When new disclosure tools or risk modules are introduced, platforms and regulators should publish evaluation summaries that report effects on recognition and behavior. Iterative evidence-based development will maximize outcomes and minimize unintended consequences (Appel et al., 2020; Thorson et al., 2021).

Implementation Roadmap

Near-term actions

Regulators should publish finance-specific examples of compliant and non-compliant posts, with annotated screenshots and model language for disclosures and risks. Platforms should add finance toggles in upload flows, require standardized paid partnership tags, and deploy accuracy prompts for posts with specific product calls. Firms should inventory creator partnerships, standardize contracts, and roll out training. Creators should update disclosure practices and produce a public statement of ethics aligned with the research on authenticity and trust (Boerman, 2017; Kim & Kim, 2021; Kay et al., 2020; Reinikainen et al., 2020).

Medium-term actions

Cross-border collaboration should establish a common core of finance promotion standards supported by interoperable ad libraries. Platforms should expand researcher access and test algorithmic demotion of content with policy violations. Educators and civil society should scale

creator partnerships for financial literacy and conduct randomized effectiveness evaluations. Researchers should report causal studies of disclosure effectiveness and the price dynamics of creator-driven attention (Thorson et al., 2021; Barber et al., 2022).

On the other hand, automated risk scoring and demotion of content should be accompanied by explanation and appeal mechanisms. Contestability of platforms actions from financial creators and transparency of the mechanisms used to discipline content is still necessary for preserving the value of genuinely educational content. Monitoring and risk-scoring systems for finfluencer content should remain assistive rather than fully automated. Creators and firms should have access to information about how they have been classified and clear criteria that can be used as reference to guide their activity without fearing sudden penalizations. Transparency and contestability are essential to avoid ambiguous repercussion on content that would result legitimate at a closer look.

Longer-term goals

The ecosystem should converge on norm bundles where creators treat finance education as a stewardship role, firms treat creator partnerships as extensions of regulated communication, platforms treat finance as a sensitive category with tailored guardrails, and regulators treat creators as part of the distribution chain with commensurate responsibilities and support (Armour, 2021; Gomber et al., 2017). Success would be reflected in durable improvements in saving and diversification, reductions in harm from speculative cycles, and reduced enforcement actions for misleading promotions (OECD, 2021; ESMA, 2024).

The recommendations here draw directly from the literature on authenticity and persuasion, the behavioral evidence on attention and retail trading, and the evolving regulatory consensus on social media promotions in finance. Implementation that privileges clarity, balance, and accountability can sustain the benefits of creator-led financial communication while mitigating the risks that arise when entertainment and markets collide (Campbell et al., 2019; Boerman, 2020; Barber et al., 2022; Cookson et al., 2023).

The literature across communication studies, behavioral finance, and regulatory insight shows that financial influencers on communication have gained resilience as agents in the modern financial communication (Symbiosis & Gandhi, 2024; Singh et al., 2025; Lai, 2025). As creators learned to perform authenticity and expertise through persona-driven storytelling, audiences responded with parasocial trust and sustained engagement, which in turn amplified reach through algorithmic ranking (Abidin, 2016; Ki et al., 2020; Appel et al., 2020; Jin et al., 2019; Reinikainen et al., 2020; Rietveld et al., 2020). This communicative strength is the source of both public value and heightened risk. From a value perspective, creators lower the intimidation barrier by turning dense financial ideas into plain language and step-by-step walk-throughs that make early moves in saving and investing feel within reach. This role aligns with evidence on message value, trust in influencer communication, and the foundations of consumer finance education (Lou et al., 2019; OECD, 2021; Singh et al., 2025). The net effect is positive when content stays anchored in literacy building and encourages diversified investing.

Risks emerge when incentives and product profiles shift toward speculative assets, high leverage strategies, or opaque offerings (Mason & Clarke, 2025). Empirical finance studies document that

attention shocks in social feeds are associated with concentrated retail flows, temporary price pressure, and subsequent reversals, patterns consistent with attention-induced trading rather than durable information discovery (Barber et al., 2022; Cookson et al., 2023; Warkulat & Pelster, 2024). Work on crypto promotion further shows that creator announcements and endorsements correlate with abnormal activity and short window price effects that do not necessarily persist, exposing late followers to potential losses (Ante, 2023; Merkley et al., 2024; Xu & Pratt, 2018).

Limitations of the current evidence point to a forward agenda. Causal identification remains challenging outside experiments that leverage platform or policy changes. Non-English ecosystems are underrepresented despite substantial activity, and long-horizon outcomes such as portfolio resilience or savings persistence deserve more attention than short-horizon trading responses (Zhang & Gong, 2021; Hudders et al., 2021). Disclosure science specific to finance would benefit from large-scale tests of placement and form in video and story contexts, guided by what is known from influencer marketing but tuned to the higher stakes of financial decisions (Boerman & Van Reijmersdal, 2020; Shan et al., 2020; Kim & Kim, 2021; Gerritsen & de Regt, 2025).

In practical terms, the path forward is not suppression but stewardship. Regulators, platforms, firms, and creators can align on a norm bundle that preserves creator-led finance's pedagogical advantages while reducing harm (Hasan et al., 2020). Clear categorical thresholds for what counts as promotion, standardized and prominent risk and compensation disclosures, partnership governance that assigns responsibility, and platform tools that encourage accuracy and reduce amplification of risky cues together constitute a feasible package (SEC, 2022; ESMA, 2024; Appel et al., 2020; Thorson et al., 2021).

TOOLS

The following sections are dedicated to introducing and describing the technical tools necessary to implement the analysis presented in the last sections of this work.

The Use of Web Crawling and Web Scraping in Academic Research

The growing massiveness of digital information has redefined how researchers find, handle, and interpret data, making web crawling and scraping some crucial methods in modern-day academic practices. Web crawling is the automated identification and information indexing of web pages, usually referring to systematically exploring hyperlinks between web domains (Olston & Najork, 2010). Web Scraping, in contrast, includes the programmed scavenging of structured/unstructured) data on web pages to process it further (Dogucu & Çetinkaya-Rundel, 2020). The two practices enable the researchers to access publicly available information and offer nascent research practices in particular research areas such as computer science, digital humanities, sociology, economics, epidemiology, and political science. However, when merged into academic research, such approaches are exposed to technical, ethical, and legal concerns.

This literature review seeks to summarize current academic views on the application of web crawling and scraping in the research environment, especially the significance to methodology, history of adoption, and ethical concerns and avenues to ethical implementation. Although this body of scholarship has been tackled technically in the past, there is an urgency to fit this into the larger academic discourse of research ethics, privacy, and compliance with the law. Bringing together knowledge bases to bridge the explanatory gap, this review will attempt to shed light on the dangers that such practices present and, as such, contribute to the arguments about responsible

digital scholarship. This review focuses narrowly on academic uses of web crawling and scraping. These activities can be utilized in many different ways in academia, including capping off bibliometric figures (Thelwall, 2008), social media (Stieglitz et al., 2018), collection of internet news archives (Rogers, 2013), and supporting the design of a more extensive understanding of computational linguistics (Roziewski & Kozlowski, 2021). These applications point to a growing reliance on academic knowledge generation on automated methods of web data gathering. However, they also create fears of data ownership, agreement to terms, and consistency with the legal regulations, like the General Data Protection Regulation (GDPR) in the European Union (IAPP, 2024). This review, hence, does not solely emphasize technical effectiveness, but also the ethical and regulatory environments within which scholarly researchers carry out their activities.

Defining Web Crawling

Web crawling is the automated, orderly navigation of links to find and index web resources. A crawler (also referred to as a spider or bot) begins with a list of seed URLs and recursively clicks on links in the page to build a graph structure of pages that connect (Olston & Najork, 2010). It is the basis of such search engines as Google and Bing, which seed billions of pages by deploying massive crawlers (Meusel et al., 2014). A standard crawler implementation will have the following components: Frontier: a list of URLs to visit, a Fetching mechanism: a way to extract HTML content, a Parser: extract hyperlinks and metadata, and a Scheduler: a way to prioritize the traversal (Kumar et al., 2014).

In academic studies, web crawling is crucial in mapping online worlds and gathering high-volume data. It can also potentially conduct analyses based on the field of study. For example, digital

humanities researchers can use crawling to recreate the old archive of news (Rogers, 2013), and computational social scientists can leverage crawling to gather information in blogs or forums, or in specialized repositories to collect data (Jungherr, 2015). In bibliometrics, crawlers can download metadata from academic repositories, like arXiv, PubMed, or the institutional open-access archives (Thelwall, 2008). Notably, most contemporary crawlers are coded to observe site-specific bans using robots.txt, which identifies legitimate lines that can be traversed by automated agents (Chang & He, 2025). Though technically a voluntary protocol, compliance with robots.txt is commonly deemed part of "polite crawling" practices.

Crawling has advanced to take advantage of the application programming interfaces (APIs) that offer data web access in a structured, rate-capped, and legally approved manner (Woody et al., 2020). As an example, the Academic Research API of Twitter has already allowed conducting massive studies focusing on political communications, misinformation, and responses to crises (Álvarez-Peralta et al., 2023). Nonetheless, API-based crawling differs from an open web crawl because platform policies and technical limits usually constrain it and are frequently subject to provider approval (Khder, 2021). It is a more extensive scale service founded on custom web scraping requests (Mitchell, 2018). It can give institutions structured data, whereby it can gain access to scraped data of their clients through API. Thus, researchers experience a dilemma between the ethical but narrow API-driven accessibility and the far-reaching, ethically questionable direct crawling.

Defining Web Scraping

Crawling is also closely related to web scraping, although these terms differ. It means explicitly extracting structured or unstructured data from web pages (Lotfi et al., 2021). Scrapers differ from crawlers in that, whereas crawlers are used mainly in navigation and indexing, scrapers are used to gather specific content like text, images, tables, or metadata. The scraping methods include HTML parsing, Document Object Model (DOM) tree-based analysis, and pattern recognition/regular expressions to isolate desired information (Mitchell, 2018). Scraping may be done manually, in which case the researcher would handcraft scripts to retrieve data from a small collection of web pages, or automatically, with tools that can operate over thousands of pages. Automated scraping has been used in the academic sphere, including: mapping publications on the news coverage during elections (Boumans & Trilling, 2016), monitoring government portals to extract public health data (Salathé et al., 2012), and mining bibliographic metadata of academic journals (Edelmann et al., 2020). Notably, web scraping, in many cases, needs to deal with dynamic web pages created by JavaScript, and a traditional scraper cannot parse these. To cope with this, systems like Selenium can be used to automate headless browsers, so the researchers can engage with websites programmatically and operate as a real user (Costa et al., 2016). Crawling is oriented toward breadth of coverage, whereas scraping is oriented toward depth of extraction. The two approaches are often combined in large-scale research projects where crawlers find and traverse the content, and scrapers retrieve particular data.

Tools and Frameworks for Crawling and Scraping

Scrapy is one of the most used open-source frameworks among various tools used to accomplish automated data collection. Scrapy is written in Python and provides crawling and scraping capabilities in one architecture. It allows the scheduling of URLs, parallel downloading, and data pipelines; thus, researchers can compile, clean, and organize data effectively (Lotfi et al., 2021). The scalability of Scrapy fits the needs of projects demanding large amounts of data on heterogeneous websites, like mapping digital news ecosystems or consolidating data on e-commerce websites to conduct economic analysis. It can also be customized thanks to its modular nature and help learn institutions where research needs demand custom interventions.

Conversely, BeautifulSoup entails a tiny HTML scraping package that works better on minor projects. Rather than offering a complete crawling infrastructure, it collects the data on a specific web page by navigating tags, attributes, and trees (Ayat Abodayeh et al., 2023). The ease has gained traction in the social sciences and humanities, where researchers prefer high ease-of-use and interpretability over large-scale applicability. It is mainly applied in educational forms as a prelude to scraping because it is an easy introduction to respondents with little knowledge in the field of programming.

In more advanced web settings, especially those that require JavaScript, Selenium gives browser control that automates the activity of a real user. Compared to Scrapy or BeautifulSoup, which parse static HTML, Selenium functions to render pages dynamically and allows scraping content created by scripts or locked behind the input form. Scholarly work exploring how to design online advertising, online platforms, or how algorithms might learn to personalize is frequently supported

by the capability of Selenium to simulate user behavior. That makes it essential to study phenomena like digital labor or interactive media whose analysis through static scraping techniques would fail.

In addition to these general-purpose systems, several application-specific frameworks have been developed in academic settings. A large-scale search engine by the Internet Archive, Heritrix, has been successfully used and applied in digital humanities and library science (Plachouras et al., 2014). Equally, services, such as Import.io or Octoparse, entail little to no-code interfaces, which reduce the technical threshold of access among researchers in fields like sociology or political science. Publish or Perish and Scholar.py are more generally used in bibliometrics to analyze the knowledge networks and research impact using citation databases in scholarly databases. Lastly, resources like Common Crawl have been built using large-scale scraping pipelines in computational linguistics and are commonly used in research involving natural language processing and machine learning (Roziewski & Kozłowski, 2021). Combined, these tools demonstrate the breadth of technical solutions to web-based research, both in beginner educational applications and in highly advanced, resource-intensive facilities.

Historical Development and Adoption in Academia

The academic history of web crawling and scraping lies in information retrieval (IR) and computer science research of the 1990s and early 2000s, when the boom of the World Wide Web generated a technical challenge and a scholarly opportunity. The early focus was on crawlers that would methodically index the ever-growing universe of pages on the web, a requirement to power the work of early search engines like Lycos, AltaVista, and subsequently Google (Brin & Page, 1998; Cho & Garcia-Molina, 2002). Scientists during this period were mainly interested in scalability,

efficiency, and coverage, and algorithms have been designed to manage the frontiers of URL spaces, duplicate detection, and link analysis (Heydon & Najork, 1999). Crawling was considered a fundamental IR technology, allowing search engines to scale from manual directories to automated search mechanisms (Arasu et al., 2001). These academic efforts can be traced in heavily referenced technical papers that codified crawler architecture and supplied mathematical models to explore traversal efficiency (Najork & Wiener, 2001). Such developments did not stay within the industry. Still, they moved into the field of scholarly practice, with researchers gradually repurposing crawlers to metadata harvesting to support bibliometric analysis, and to crawl web corpora to support natural language processing (Thelwall, 2001; Gulli & Signorini, 2005).

A similar technique, scraping, emerged during the period in parallel to crawling, but was not initially differentiated into a distinct research practice. Scraping was treated as the secondary method, which aimed to turn what was retrieved in HTML into structured datasets that could be further analyzed (Laender et al., 2002). Scraping tools in their early days were primitive in terms of modern frameworks, usually based on regular expressions or customized hand-coded parsers. Although simplistic, these tools allowed innovative academic applications like the ability to construct surveys through the web, domain-specific search engines, and build specialized repositories (Duka et al., 2023).

Therefore, the 1990s-2000s were the initial years of the technological grounding of crawling and scraping as scholarly approaches. Whereas crawling was celebrated as having made the infrastructural contributions to enable the large-scale indexing, scraping was aligned as a tactical instrument that scholars could use to convert the disorganized content into formats they could

analyze. Initially, these practices were the domain of computer science, IR, and computational linguistics, but they preconditioned their later extension to social sciences and humanities.

Expansion to Social Sciences

In the late 2000s and early 2010s, crawling and scraping became increasingly used within the social sciences as social media, online news publications, and government databases were created. Scholars realized that the web was not just a technical artifact anymore but an active social world, whose behavior, culture, and politics generated large amounts of data. Such a transformation expanded the use of scraping beyond infrastructure construction to empirical social inquiry. The study of Twitter and alternative social media networks was among the first and most impactful fields. By scraping, social scientists targeted the collection of tweets, retweets, follower networks, and generated datasets, making it possible to perform analyses of political communication, social movements, and crisis response (Java et al., 2007; Bruns & Burgess, 2011; Stieglitz et al., 2018). The Twitter platform was particularly popular due to its comparative openness, and the early release of APIs enabled automatic data acquisition (Tufekci, 2014). Such works led to unique and unexplored revelations about mass community speech, yet generated methodological tension over representativeness, sampling biases, and ethical concerns of gathering user-generated content without express permission (Zimmer, 2010; Fiesler & Proferes, 2018).

In addition to social media, scraping was adapted to track the news media, forums, and government websites. As illustrations, Boumans & Trilling (2016) invented a technique to scrape digital news archives to track media coverage on elections. Salathé et al. (2012) scraped official web pages to obtain epidemiological data on public health during a crisis. Through scraping, Jungherr et al.

(2015) explored online political campaigns in political science and were rebuilding digital news ecologies. Its prevalence in these settings indicated its potential to offer real-time, large-volume data, as researchers could notice the phenomenon at magnitudes that would initially have been unachievable via conventional survey or interview procedures. Notably, this growth did not come without criticism. Researchers pointed out that scraping publicly available information, such as social media posts, attracted the risks of privacy infringement, terms of service violation, and questionable ethics (Khder, 2021). These controversies marked a shift in literature, moving beyond technical considerations into normative claims. This theme is now central to any scholarly appraisal of web data approaches.

Current Uses Across Disciplines

Web crawling and scraping have become common research practice in various academic fields today. They are applied far beyond computer science and the social sciences, to digital humanities, bibliometrics, business analysis, and epidemiology. Crawling and scraping have allowed the digital humanities to conduct projects, including large-scale textual mining of historical archives, recovering lost or fragmented web domains, and analyzing cultural phenomena at scale (Rogers, 2013). For example, researchers have crawled and archived historical websites, using Heritrix and the Wayback Machine provided by the Internet Archive (Weigle, 2023), and scraping pipelines have been employed to assemble datasets to support literary and cultural analysis. Such projects highlight that automated data collection has become a scientifically inquisitive and cultural heritage preservation tool.

In addition, scrape utilities like Publish or Perish and Scholar.py have become common in bibliometrics and scientometrics practices, to scrape citation data across the Google Scholar, Scopus, and Web of Science platforms (Bornmann et al., 2016). They have facilitated simple research studies on research impact, knowledge networks, and scientific collaboration. Institutional repository scraping has also been applied to promote open science movements that support accessibility and transparency of the research studies (Ahmed & Othman, 2021). There is also embedded scraping in business and market analysis, where there are analyses of online labor markets, online price examinations of e-commerce strategies, and printer customer behavior. The business and market analysis sphere has also adopted scraping to examine e-commerce pricing, online labor markets, and consumer trends. As one of the examples, Amazon (or eBay) scraping with scrapers allowed researchers to understand dynamic pricing strategies, high-velocity competitive strategy, and consumer trends (Guyt et al., 2024). In the same vein, labor economists have used scraping techniques to review job websites such as Upwork or Fiverr to provide an empirical account of the changes occurring in digital work (Green et al., 2018). Scraping in epidemiology and public health has been significant in studying disease outbreaks and health behaviors. Salathé et al. (2012) have exemplified how scraping can be applied to obtain real-time epidemiological statistics about government portals and health forums. In a more recent example, scraping was invaluable in the case of the COVID-19 pandemic to gather information on the number of cases, public policy actions, and misinformation that flowed to not only make decisions in real-time situations but also retrospective analyses (Chen et al., 2020; Raamkumar et al., 2020). In combination, these applications of disciplines demonstrate how crawling and scraping have become normalized academic activities. They are no longer stuck in technical ghettoes but are the methodological mainstays of fields as divergent as sociology, linguistics, economics, and medicine.

Trends in Publication Growth

The increased academic dependence on web crawling and scraping is perhaps best illustrated by publication trends, highlighting the rate at which the two methods have adopted a standard in scholarly practice. A once niche group of methods in computer science, over the last 20 years, has grown into an interdisciplinary methodological baseline. The indicator in systematic reviews and bibliometric surveys consistently suggests that research involving crawling and scraping is exponentially increasing, signaling the shift of technical novelty towards the mainstream (Khder, 2021). According to a thorough literature review by Khder (2021), the volume of publications based on web data collection almost tripled between 2010 and 2020, with particularly steep increasing rates in the social sciences and health spheres. Computer science and information systems remain significant areas of origin. Still, since then, the highest growth has been in digital humanities, business analytics, and public health, which were the last to embrace the methods but now see them as essential. (Dogucu & Çetinkaya-Rundel, 2020) also refer to normalization of scraping by introducing the practice into data science schemes, where scraping is now introduced alongside statistical analysis and machine learning.

This change is vividly explained by quantitative evidence. Bibliometric studies of journals on computational social science and digital humanities show that the number of articles on scraping has sharply increased, especially since 2015 (Stieglitz et al., 2018). Specifically, Cheng et al. (2024) describe how the percentage of peer-reviewed articles where web scraping techniques are indicated in social science journals rose over 250% in seven years between 2015 and 2022, highlighting their fast institutionalization. Likewise, online epidemiological reviews report that scraping has become

an ordinary feature of infectious disease surveillance and the prospective tracking of individual health actions. This shift in direction can be said to be indicative of larger changes in academia: the increased focus on big data, digital trace analysis, and methods that are computationally demanding in research design (Cheng et al., 2024). Concurrently, ethical and legal debate has been compounded by this development. The higher the rates of such practices, the higher the concerns about privacy, consent, and compliance with acts like the GDPR, as Zimmer (2010) and Fiesler & Proferes (2018) note.

Ethical Challenges

The rapidly increasing practice of crawling and scraping websites as part of scholarly research has yielded methodological potential and necessary debates over questions of ethics. They comprise intellectual property rights, data ownership, prior consent, user autonomy, and the dichotomy between the legal systems and the research practice. Whereas the technical aspect of crawling and scraping has been handled mostly, the ethical aspect of the same is very questionable, as more ethics and regulatory intervention have set in, which are coupled with consumer resentment of internet privacy.

Intellectual Property and Copyright

Among the most enduring ethical arguments on web crawling and scraping in scholarly work is that they tend to violate copyright and intellectual property issues. Crawling, by definition, is the automatic reproduction of web pages to be indexed, archived, or analyzed in large quantities. Although commonly positioned as an infrastructural practice, crawling may create legal conflict where copyrighted content is copied in high volumes. For example, the Wayback Machine, the

Internet Archive crawler and web archive, has received legal opposition from publishers claiming that archiving is unauthorized duplication of copyrighted writings (Milligan, 2017). Likewise, the use of academic programs that index digital libraries to create corpora to facilitate bibliometric or linguistic research brings into question the implications of replicating the entire, unlicensed collection of articles without permission, violating the rights of publishers (Khder, 2021).

Scraping is more limited in scope but tends to raise particularly intense attention since it comprehensively harvests and reuses specific texts, photographs, or metadata. Copyright often safeguards these aspects, and converting them into table-based data can break intellectual monopoly, even when the text was open to the general Internet audience (Boyd & Crawford, 2012). Scholars sometimes use the concept of fair use to justify scraping when it can be presumed non-commercial, scholarly, and transformative in purpose, e.g., scraping news articles to generate discourse patterns (Urban & Quilter, 2016). Nevertheless, court results are not always consistent. As is illustrated in the case of *Associated Press versus Meltwater* (2013), (Justia Law, 2013), the court dismissed the argument of transformative use by Meltwater. Instead, it held that aggregating news snippets was an infringement and not a form of scholarship. The difference between crawling as infrastructural and scraping as extractive is thus both ethically and legally meaningful, but both concern the reproduction of intellectual work. Academic researchers, more broadly, have the challenge of not only acclimatizing to different regimes of law, like the more stringent copyright laws in the European Union compared to the United States, but also maintaining a sense of moral obligation towards the respect of authorship and creative work. As Geiger et al. (2018) caution, the promotion of knowledge needs to occur in a balance with protecting the intellectual contributions of others.

Data Ownership and Licensing

Data ownership and licensing lie at the heart of the ethical analyses of web crawling and scraping. When crawling is used in proprietary repositories like Scopus, JSTOR, or Web of Science specifically, there is a risk of crawling beyond the set boundaries of permitted access even without an express scraping action. Mass crawling of these databases imposes a technical burden on the providers and may also breach contractual agreements that guide institutional subscriptions (Mongeon & Paul-Hus, 2016; Beel & Gipp, 2010). The *Aaron Swartz* case shows the stakes involved: Swartz carried out automated downloading of JSTOR articles using indexing bots, posing his move as an act of intellectual emancipation. The case led to federal prosecution, leading to a debate between activism, academic inquiry, and intellectual property infringement (Samuelson, 2013).

Scraping heightens these ownership issues by converting raw internet content into organized datasets that can be redistributed or reutilized. Although services like Twitter and Facebook offer API licenses to control scholarly access, researchers often bypass the services by simply scraping the web (Zimmer, 2010). These practices are incompatible with licensing terms and can weaken fragile trust between academia and data providers. In one example, Elsevier and other large publishers have made multiple notifications attempting to halt illegal crawling and scraping of their journal platforms, highlighting their belief that scholarly articles are proprietary rather than open sources (Bodó et al., 2018). Khder (2021) claims ownership disputes reveal a bigger clash between the "data capitalism" system and open science. Corporation commercializes digital trails and limit access to knowledge with restrictive licenses; academia, in contrast, celebrates a culture of openness, reproducibility, and democratizing knowledge on the one hand. To researchers, the

question is no longer whether it is technically possible to see the material, but how it is legally, ethically, and socially acceptable.

Informed Consent and User Autonomy

Research ethics has long hinged on the need to gain informed consent, yet using informed consent in web crawling and scraping leaves much to be desired. The crawlers often visit personal blog pages, online forums, or community archives to harvest publicly visible content. This material is technically in the public realm. Still, the people making the posting do not usually expect such material to be stored in comprehensive academic databases or interrogated outside its targeted audience. Mechanisms, like the robots.txt protocol, to signal that crawling is allowed or denied, are an effort to enforce autonomy, but doing so is not mandatory and is not always respected. Most researchers ignore such limitations because the information has a high academic worth (Chang & He, 2025).

These concerns are enhanced through scraping since, in many cases, they scrape off personal or behavioral information on social media sites and interactive sites. The Harvard Tastes, Ties, and Time (T3) project, which scraped the Facebook profiles of the undergrads without their consent, exemplifies how they can destroy user autonomy as well as elicit reactions that backfire when datasets are later released publicly (Parry, 2011; Zimmer, 2010). Although information may be publicly visible, its repurposing into research would violate the notion of contextual integrity developed by Nissenbaum (2010): no one should assume that information shared in one social setting should be freely transferred to other settings.

Further complicating the issue is that most users are unaware of terms of service (ToS), data-sharing policies, or even the technical avenues through which their information can be harvested. Fiesler & Proferes (2018) believe this level of information asymmetry makes any conventional understanding of the concept of consent, in online conditions, mostly elusive. To researchers, conflating public accessibility with ethical use can harm the autonomy and expectations of people whose data is reused. Crawling and scraping both necessitate a reevaluation of how informed consent is supposed to work in a digital research setting: not only is it required to be technically implemented, but the rights and intentions of users must be approached ethically as well.

Website Terms of Service (ToS) Conflicts

Terms of Service (ToS) agreements are one of the key areas of ethical conflict regarding web crawling and scraping in scholarly research. Websites commonly limit automatic access with specific contractual provisions or by an automated mechanism such as robots.txt. Crawlers often block harvesting large pages to preserve server bandwidth, and structured data scraping is explicitly stated when scraping or circumventing licensed APIs. Despite these limitations, research groups often use automated techniques to analyze phenomena that interest the community. For example, researchers have scraped government websites and health portals to gain timely information about disease outbreaks and policy initiatives during a public health crisis- often in direct violation of ToS with crawlers or scrapers (Salathé et al., 2012).

Legal conflicts support the unpredictability of ToS application. This tension was exemplified in the *hiQ Labs versus LinkedIn* case (2017-2022): LinkedIn invoked ToS and stated that scraping constituted Computer Fraud and Abuse Act (CFAA) violations, and courts said that public profile

data was not subject to an unlimited fencing off through contractual terms (Justia Law, 2022). Despite allowing certain leeway regarding access to public information, this judgment did not establish whether academic research is treated similarly, especially in jurisdictions that filed stronger interpretations of the contract (Park, 2025).

ToS disputes regarding the confrontation between corporate authority and academic accountability are cast in the scholarly discourse. Some claim that adherence to the strict approach essentially transfers control over the data of public interest into the sphere of the private platforms, thus limiting the freedom of inquiry and strengthening the monopolies of data distribution (Fairfield & Engel, 2015). Another side warns that violating ToS may lead to reputation loss, legal reprisal, and the loss of social confidence in research enterprises. ToS disagreements thus raise an even further philosophical question: Would it be better for scholars to follow the contractual dictates of corporations or the normative duty of producing knowledge beneficial to society?

Risks of Re-Identification in Scraped and Crawled Data

The possibility of re-identification of anonymized datasets crawled or scraped is one of the most urgent ethical concerns, even when these datasets are anonymized. Depending on the crawler, it often captures all the web pages, such as usernames, time stamps, and contextual information that can be reassembled to reverse engineer unique identities. Datasets obtained by scraping, where the data is compiled on Twitter, Reddit, or health forums, might seem anonymous when lacking any explicit identifiers, but can be linked back to other datasets to de-anonymize people (Narayanan & Shmatikov, 2008; De Montjoye et al., 2013).

Such risks become especially sharp when vulnerable groups are involved in research. As an example, crawled cancer patient support sites, or scraped personal blogs on mental illness, may implicitly disclose a health condition, sexual orientation, or political membership and therefore result in stigmatization or discrimination (Eysenbach & Till, 2001; Mahoney et al., 2022). Notably, people who post such data online do not usually expect that these data will be archived, processed, and potentially de-anonymized in research endeavors. Khder (2021) stresses how increasing the artificial intelligence of machines and pattern recognition compounds such risks. A data set that is safely no longer identifiable today can once again be identifiable tomorrow because algorithms are more sophisticated. Due to this fact, scientists should be cautious about their methods: they need to collect the least amount of sensitive variables and anonymize them with the best current methods of doing so; they also should weigh not only short-term risks but also how collected data can be used far in the future.

Case Studies of Ethical Controversies

Ethical dimensions of crawling and scraping are most evident when considered in more high-profile controversies that have informed scholarly and public discussion. The incident involving Cambridge Analytica is the most common example of misusing data in the name of research. The political consulting company gained access to the personal information of over 80 million Facebook users with the help of a personality quiz application initially presented as an academic initiative (Cadwalladr & Graham-Harrison, 2018). The users had signed a consent form for the app's data collection. Still, they did not expect their data and Facebook friends' data to be mined and used in political microtargeting. The case represents the breakdown of informed consent and contextual integrity from an ethical perspective (Nissenbaum, 2010). Users believed their posts and

preferences would be utilized socially and not in mass political profiling. To academia, the scandal reaffirmed the reputational risk of techniques considered scraping or comprehensive crawling of personal information. Although Cambridge Analytica was not a part of the academic establishment, its connection to research raised doubts about whether academic data practices and business abuse could ever be decoupled.

The example of the internet activist and programmer Aaron Swartz further demonstrates ownership and licensing tensions. Swartz employed an automated crawler to look through JSTOR archives, downloaded millions of articles through the MIT network, and justified his actions as part of a free knowledge campaign (Samuelson, 2013). He was charged with federal offenses of wire fraud and computer abuse, with the harshness of the persecution also coming in for much criticism after he died in 2013. Swartz was involved in activities that broadened the boundaries between activism, scholarship, and piracy. On the one hand, his crawl revealed the limitations of the licensing regimes that prevent open science; on the other hand, it posed questions about contract breach and the burden on the infrastructure of databases. To the academic researchers, the case raises the question of whether civil disobedience in the name of the open access policy can be reconciled with institutional ethics.

Collectively, these ethics issues show that crawling and scraping lie in a gray area between creativity and copyright violation, transparency and blockage, research and attacks. Intellectual property, licensing issues, informed consent, ToS, and re-identification risk demonstrate that knowledge may conflict with the rights of persons, companies, and societies. The cases presented show how an ethical violation in glaring omissions, activism, or profit-seeking motives can easily

go into the territory of a trust and legality crisis and cause lasting effects on academic research. To scholars, the lesson is not one of obedience but one of developing a principled way: evaluating the possible merits of research on the one hand against the loss of autonomy, privacy, and trust on the other. With the change in digital settings and the increase in regulatory bodies, the field of ethical responsibility is not an option to take lightly.

Conditions for Ethical Viability

Web crawling and scraping have emerged as tools in academic research. Still, their ethical validity lies not merely in the fact that such practices are technically feasible, but that such activities can be performed to advance professional ethics standards and the law. Although discussions of copyright, ownership, and autonomy will continue to dominate, two criteria have become particularly compelling toward determining ethical validity. Adopting the polite crawling regulations as a minimum technical-ethical specification and aligning data collection and processing activities with the General Data Protection Regulation (GDPR) within the European Union.

Polite Crawling as an Ethical Baseline

Polite crawling has been identified as one of the criteria for evaluating the ethical feasibility of auto-scraping data in academia. In contrast with aggressive or mindless crawling, which cares less about site integrity, polite crawling is technical, restrained, procedural, transparent, and convention-based. It includes constraints like rate limiting, following of robots.txt control guidelines, a definite indicator of crawler agent, and overall damage reduction (Chang & He, 2025). In this context, polite crawling is more than a technical best practice; it is also a normative ideal of responsible digital scholarship.

The most prominent aspect of polite crawling may be the robots' exclusion protocol (robots.txt). Robots.txt is one of those gatekeeping mechanisms that allows site administrators to mark which portions of their websites can or cannot be used by crawlers, indicating the preferences of content owners in that regard. Though robots.txt has no legal force, compliance remains an ethically mandated minimal requirement (Karwatzki et al., 2017). Academic researchers, in particular, insist on compliance as a sign of recognition of the autonomy of the platform and a way of minimizing reputational risks posed by the perceived intrusiveness or exploitation of a project.

However, the protocol is limited. Researchers like Fiesler & Proferes (2018) believe that a decision to use robots.txt is not a decision that the interest of web administrators should apply, but of the knowledgeable resolutions of particular users whose details are stored in these segments. This forms a paradox because, although this polite crawling is respectful of platform-level orders, it does not always raise the deeper issue regarding user consent or contextual integrity (Nissenbaum, 2010). Therefore, robots.txt is not synonymous with complete ethics compliance since it provides the minimum protection.

The concept of rate limiting is also at the heart of polite crawling; it prevents overloading of servers and the inconvenience to users caused by an excessive number of automated requests. Uncontrolled crawling, especially on an industrial scale, may result in site performance degradation, resulting in denial-of-service effects. Ethically, this qualifies as a breach of the principle of non-maleficence that requires a researcher not to act in a manner that causes harm (Floridi & Taddeo, 2016). Rate limiting is a form of technical protection and moral compulsion. Academic crawlers may

appropriately tune the frequency of requests to balance the interests of obtaining rich datasets and maintaining the functionality of sites. This is ensured by the principle of data minimization discussed in GDPR and the design and conduct of traditional research, which stipulates that researchers acquire only the data necessary to state well-defined scholarly goals (Tsvetkova et al., 2017). Not only does this mitigate infrastructural harm, but it also curbs the risk of excessive data collection of personal information.

Polite crawling also requires openness in crawler identity and approach. Technical implementations include a user-agent string that identifies the crawler and, preferably, a contact email address. Although a relatively minor consideration in commercial scraping, the transparency of crawler design is particularly critical to academic use cases due to the necessity of accountability to institutions, funders, and the general population. The professional codes of conduct justify this focus on openness. The ACM Code of Ethics 2018 explicitly demands that computing professionals uphold honesty, transparency, and accountability in their practice. Gogoll et al. (2021) point out that although codes such as the one of the ACM are undoubtedly imperative, deliberation is an essential accompaniment to them, as with web crawling, where it is highly valuable to have transparency and accountability. On the same note, the APA Ethics Code (2017) also notes that clear communication and abstinence from deception are crucial in any research setting. To the digital scholar, this means publicly recording crawl parameters, stating whether robots.txt and rate limiting were detected, and revealing how the resultant data will be used. These protective measures create trust and reinforce defensibility in case of legal or reputational attack on the research.

Although general polite crawling may establish the vital technical and procedural responsibility benchmark, it is not enough to address broader ethical and legal issues. In particular, it fails to solve the problems of collecting personal information, the dangers of re-identifying an individual, and whether an overall regime of privacy protection like the GDPR is met. That is, polite crawling guarantees that the data collection instruments are as non-invasive as possible and non-arbitrary to the platforms they parse. Yet, it does not in itself conclude that the purposes of data use are not misaligned with the ideals of user autonomy, fairness, or legal conformity.

GDPR and the Regulation of Research Data

The legal regime with the most thorough take on the processing of personal data is the General Data Protection Regulation (GDPR), which came into effect in the European Union in 2018. It is not limited to commercial actors, but also considers academic researchers who perform web crawling and scraping. According to Mondschein & Monda (2018), GDPR has fundamentally changed how ethically and legally sound digital research has to be conducted by setting specific standards that harmonize the relevance of individual rights and data accessibility. To a researcher, this indicates that the practices of crawling and scraping should not only be considered as far as their technical advantage or disadvantage applies, but they should also be considered, given the regulatory schemes of protection of personal privacy.

Definition of Personal Data

GDPR is applicable where data is personal data that is widely understood as any information that relates to an identifiable, natural person. This encompasses both direct identifiers, e.g., names, email addresses, or IP addresses, and indirect identifiers, which, when used together, can be used

to re-identify an individual (Article 4(1), GDPR). Recital 26 clarifies that pseudonymized or publicly available data may be classified as personal data in case people can still be recognized. This has direct implications for crawling and scraping: whereas websites such as Twitter or Internet forums may give data publicly, the combination of posts, screen names, or post times into large datasets can allow re-identification, both when re-identified with other sources (De Montjoye et al., 2013). Accordingly, academics are not free to make the presumption that something public is ethical or legal.

Lawful Basis for Processing

In GDPR, any processing of personal data must be based on one of six lawful grounds (Article 6). There are three of them that are most relevant in an academic setting.

1. Consent- explicit participant consent, which is not always feasible on large-scale crawling or scraping due to millions of users involved.
2. Legitimate Interests- the processing is fine when the research can serve a higher societal good beyond the risks involved to the studied individuals. Again, this must be performed with serious balancing tests.
3. Public/Scientific Research- Article 89 enables limited derogations to research, allowing some flexibility (e.g., to prolong the storage or make secondary use) subject to some safeguards (e.g., pseudonyms or encryption).

Ethical soundness thus relies upon the careful definition of the legal foundation of the project, presenting the report of its necessity, and explaining proportionality between the overall well-being of the society and the risk to an individual (Mondschein & Monda, 2018).

Core GDPR Principles

Several GDPR principles determine the scope of crawling and scraping:

1. Purpose Limitation Art. 5(1)(b): Data should not be gathered without clear research objectives and cannot be used repeatedly.
2. Data Minimization (Art. 5(1)(c)): Only the data that will be needed to answer the research question is to be gathered.
3. Storage Limitation (Art. 5(1)(e): Information must not be stored longer than required to achieve the research goals.
4. Integrity and Confidentiality (Art. 5(1)(f)): The datasets should be shielded by adequate technical and organizational safeguards against breach or other unauthorized access.
5. Accountability (Art. 5(2)): Institutions and researchers must show proof of compliance, which can usually be done by working in conjunction with Data Protection Officers (DPOs) and Institutional Review Boards (IRBs).

Combined, polite crawling and GDPR compliance set the dual contexts in which web crawling and scraping can become ethically acceptable in academic research. Polite crawling acts as a technical and procedural base, guaranteeing that the data collection process causes minimum harm, does not discriminate against anyone who operates sites, and is transparent. GDPR, in turn, offers the normative and legal framework, which focuses on individual rights, illuminates legal grounds on which processing is possible, and imposes accountability. Nothing is adequate on its own; a crawler can constrain itself to honor robots.txt but abuse privacy by mishandling personal information, and simply following the rules of GDPR without a technical bridle can still overload the online systems.

Ethical feasibility, hence, depends on their synthesis. Researchers will have to integrate technical courtesy into legal protection, but they will also have to consider the emerging user expectations of privacy. This combination enables a responsible development of academic inquiry that does not recede into mistrust in digital research.

Toward an Ethical, Privacy-Respecting Approach

The above discussions indicate that although crawling and scraping represent a tremendous methodological potential for academic research, the ethics and legality of these actions are highly controversial. The issue then lies not in whether such practices are permissible, but in the conditions under which such practices may be carried out with the understanding that they will, in some way, promote knowledge and ensure privacy, intellectual property, and the user's autonomy. Basing ideas on the technical literature on polite crawling, the legal regulation provided by GDPR and CCPA, and the ethical statement of such professional associations as the ACM and APA, this section will distill a style of crawling and scraping that can be said to be moral and respectful of privacy within the context of academic research.

Technical Safeguards: Polite Crawling as the Operational Foundation

On the technical level, any successful strategy should start with the base of polite crawling. Polite crawling involves rate limitation, obeying the robots.txt exclusionary guidelines, proper identification of the user-agent, and minimization of the load on the server (Kolobov et al., 2019). This means data must not compromise the integrity of visited sites, and that the researcher does not adopt deceptive means in data collection. To conduct research, these safeguards must be enhanced beyond what is required by observing more stringent technical and procedural measures. The rate

limiting must be adjusted to the recommendations, substantially lower than rates that could produce service degradation, and the crawl schedules need to explicitly not coincide with usage peaks. Similarly, adherence to robots.txt ought to be positioned as an obligation, despite the lack of any binding force, as it is an expression of stated desires of content providers and a demonstration of respect to their autonomy. Identifying the crawler differences using the user-agent strings and the contact information adds more transparency and accountability, where the site operator can query or dispute the crawl. Further, data minimization would not just be limited to the number of requests made but also to the scope of information being gathered. In line with GDPR and research ethics, this principle makes researchers think about collecting only the essential information to respond to clear scientific questions (Tsvetkova et al., 2017). Together, these create the technical restraint basis of making ethical crawling and scraping operations, and underline that the sufficiency of these measures all lies in the practical application of game-changing privacy safeguards.

Privacy-Respecting Design

Of particular importance is the processing of personal data. Scraping and crawling constantly encounter the material that can be classified as personal data under GDPR (names, usernames, IP addresses, timestamps, or combinations that enable identification). Even publicly observable data, such as tweets or posts in a forum, remains personal so people can be identified again (Recital 26, GDPR).

A privacy-sensitive approach thus has to implement hierarchies of security:

1. Anonymization and pseudonymization: Personal identifiers must be deleted or anonymized whenever possible. Pseudonymization (replacing identifiers with codes kept in a distinct

place) could allow valuable analysis and still afford identity protection. Nevertheless, combining datasets, as demonstrated by Narayanan & Shmatikov (2008), is always associated with a risk of re-identification. Scientists should be aware of this pitfall and implement the latest approaches, like differential privacy, to minimize the threat in the long term (Dwork, 2008).

2. Purpose limitation: The data can only be utilized according to the specified research purposes, per Article 5(1) (b) GDPR. The unlimited reuse of scraped information goes against contextual integrity (Nissenbaum, 2010) and user expectations.
3. Data retention limits: This is aligned with Article 5(1)(e), which holds that scraped datasets must only be saved when a research goal is not attainable and where deletion procedures must be followed after.
4. Risk assessment for vulnerable groups: Studies based on data from health forums (or children or marginalized groups of the population) should be approached with greater attention to scrutiny. In this case, GDPR Article 9 on special categories of data is especially applicable, because it introduces the need to take additional precautions when handling the special categories of data, such as encryption, further minimization, and, when possible, consent.

Privacy protection requires a multilevel proactive approach: anonymization to protect in the short term, minimization to reduce exposure, and contextual awareness to help prevent abuse.

Legal and Ethical Frameworks

Adherence to the data protection laws gives ethical viability to scraping its normative outline. Under the GDPR, academic researchers can use several grounds: legitimate interests (Article 6), the public interest in scientific research (Article 89), or, with an additional clause, consent. Each entails proportionality between the societal price of the study and the detriments to individuals (Mondschein & Monda, 2018). On a similar note, the California Consumer Privacy Act (CCPA) grants rights to California residents to know about, opt out of, and request deletion of data collected about them. Although CCPA is not directly academic research specific, it is part of the wider cultural change where data subjects insist on control of their information. Therefore, any legitimate model would incorporate ideas of transparency, such as a clear definition of the research purposes, a contact person, and opt-out provisions.

Ethical viability is also formed by copyright and fair use. Courts have unevenly dealt with fair use in actions concerned with automated gathering (Urban & Quilter, 2016). Nevertheless, researchers tend to note that when the purpose of scraping is not-for-profit, transformative, and socially constructive, such an action is better justified than unconditional copying. This implies that academic projects will have to be focused on transformative value, e.g., making discourse analysis, monitoring of public health, or other research outcomes that are otherwise not replaceable with the original materials.

Integration into a Holistic Ethical Model

Combining these layers, technical protection and legal regulation can result in an ethically viable and privacy-protective strategy. Technical solutions like polite crawling and data minimization are effective measures against harm, but not a surety against privacy infringement. Legislation like the GDPR provides adequate protection to personal information, but they do not assume the infrastructural strain of the big-scale crawling. Ethical codes would give the researcher integrity and must be backed up technologically and legally. Implementing these two dimensions would aid the academic fraternity in developing a paradigm of web crawling and scraping that would not only be acceptable in society but also ethically permissible and socially justified. In this regard, the engagement of technical skills and legal regulation is the key to responsible digital scholarship in an automated data collection age.

This literature review discusses the role of web crawling and web scraping in performing academic research, exploring technical backgrounds, historical use, and the ethical situations they present. Since early work in information retrieval and bibliometrics, automated data have come to play a central role across the social sciences, digital humanities, and epidemiology in studying online spaces. This growth has, however, been shadowed by intricate discussions about intellectual property, licensing, informed consent, user autonomy, problems encountered with Terms of Service, and the ever-looming threat of re-identification. One major conclusion is that ethical feasibility cannot depend solely on technical feasibility. The review identifies polite crawling as a form of responsible research practice, including following robots.txt, rate limiting and data minimization. Still, technical restraint should be combined with strong legal and ethical frameworks. The GDPR gives the most robust system of requirements on lawful basis, purpose

limitation, minimization, and accountability, and additional frameworks, including the CCPA and professional codes (ACM, APA, ALLEA), support transparency, fairness, and proportionality requirements. The new path towards future-oriented research involves a multisided strategy with multifaceted but unified solutions embodying technical, legal, and institutional measures. Although there is an improvement, there are still significant gaps in adjusting research ethics standards to online conditions and introducing discipline-specific codes of practice.

Large Language Models for Data Extraction in Research

Research teams face a growing tide of unstructured text. Articles, reports, policy memos, and clinical notes carry facts that matter for science and practice. Classic extraction pipelines split work into many steps, such as tokenization, tagging, and hand-built rules. Those steps were fragile across fields and formats, and they demanded long upkeep and costly tuning (Lopez, 2009; Tkaczyk et al., 2015). Large language models change that workflow. These models read raw text and return structured fields in one step or a short chain. They can extract entities, relations, and values with prompts and light guidance (Minaee et al., 2024).

The core claim of this chapter is direct. Large language models are a valid and effective way to extract research data from unstructured text when paired with schema constraints, retrieval, and careful checks. This approach reduces manual effort, speeds synthesis, and adapts across domains with minimal code and modest hardware (Lewis et al., 2020; Khattab et al., 2023). The chapter offers background, shows where models work well, lists open models to consider, and explains why the Llama family and the Ollama framework provide a convenient path from a pilot to a stable pipeline.

A Short History of Large Language Models

Early language models counted n-grams and backed off when context ran thin. Those models could not hold long context or rare terms well. Recurrent neural networks arrived, yet struggled with long sequences and training stability because gradients either vanished or exploded. The Transformer made a clean break by focusing on attention, parallelism, and long-range links across tokens

(Vaswani et al., 2017). This design lets training scale across many devices and keeps gradients stable over long texts, which raises both quality and speed.

Subword tokenization also helped. Byte Pair Encoding and SentencePiece map rare terms into shared parts, which improves coverage for names, drugs, alloys, and novel compounds (Sennrich et al., 2016; Kudo & Richardson, 2018). Subword units lower the out-of-vocabulary rate and allow the model to copy spans with fewer errors during extraction tasks. This feature proves helpful when text contains symbols, hyphenation, or long chemical names.

Pretraining at scale followed. Autoregressive models learn to predict the next token, which teaches broad world knowledge and discourse skills from large corpora (Brown et al., 2020). Masked language models learn by filling blanks and use deep bidirectional context to capture rich relations within a window of text (Devlin et al., 2019). Instruction tuning and reinforcement learning from human feedback, then aligned models to follow prompts in plain language and to refuse some unsafe actions when asked (Ouyang et al., 2022; Christiano et al., 2017). Together, these moves produced modern large language models that are practical for extraction.

Efficiency advances matter for research groups. LoRA adds small trainable adapters to cut fine-tuning cost and preserve base model knowledge (Hu et al., 2022). QLoRA pushes memory use down by training on four-bit weights while keeping quality high for many tasks (Dettmers et al., 2023). For serving, eight-bit kernels keep throughput high on modest hardware without large accuracy losses (Dettmers et al., 2022). Retrieval augmented generation adds context from an

index, which helps models stay grounded and traceable in scholarly extraction tasks (Lewis et al., 2020).

Functioning Principles Relevant to Extraction

Tokenization shapes what the model can read and copy out. Subword units help the model match technical phrases across a field, and they let the model copy exact spans for numbers, units, and chemical names with fewer errors (Sennrich et al., 2016; Kudo & Richardson, 2018). Attention links cues across distant parts of a document, which supports document-level facts, such as matching a method and an outcome reported in different sections or tables (Vaswani et al., 2017; Jain et al., 2020). Many extraction targets depend on cross-sentence links, so the ability to hold and align evidence across distant sentences is central for many reviews.

Mixture of experts designs route tokens through subsets of weights, which raises capacity without linear cost growth. This idea can deliver high quality at useful latency when routing is stable and batching is tuned (Shazeer et al., 2017; Fedus et al., 2021). For extraction, this can matter when models must read long documents while staying responsive for human review. Adaptation lowers the barrier to domain use. LoRA and QLoRA allow quick tuning for a field and keep privacy risks low by enabling local work with limited compute budgets (Hu et al., 2022; Dettmers et al., 2023). Quantization and memory-aware kernels allow teams to run strong models on a single workstation or a small server cluster (Dettmers et al., 2022; Frantar et al., 2022). Retrieval grounds output in the source text and supports audits by making citations easy to surface. Grammar-constrained

decoding then keeps outputs in a strict schema and reduces cleanup time (Beurer-Kellner et al., 2024; Park et al., 2025).

How LLMs Are Used as a Data Extraction Method in Research

Many studies now test large language models for research data extraction. In health research, teams have used prompts to extract PICO frames, outcomes, trial arms, sample sizes, and follow-up periods from abstracts at scale. A recent proof of concept extracted more than six hundred eighty thousand PICO frames from PubMed abstracts using staged prompts and validation (Reason et al., 2024). Other studies compare model extractions to human coders for systematic reviews. Results are mixed but trend positively, especially when prompts include schema hints and when retrieval and validation steps are used to anchor each field (Schmidt et al., 2024; Sun et al., 2024). In materials science, a study introduced a staged method named ChatExtract that improved accuracy by adding targeted follow-ups for hard fields (Polak & Morgan, 2024).

Document-level extraction remains hard because facts can span sections, figures, and tables. The SciREX dataset pushed models to link questions, methods, and results across full papers, not just single sentences (Jain et al., 2020). Large corpora such as S2ORC support training and evaluation for this kind of task by providing millions of papers with parsed structure and metadata (Lo et al., 2020). Classic PDF tools still add value. GROBID and CERMINE extract sections and references, which give models cleaner inputs and better anchors for retrieval (Lopez, 2009; Tkaczyk et al., 2015).

LLM extraction is valid and effective under clear constraints. Retrieval supplies ground truth passages. Constrained decoding enforces a JSON schema. Self-consistency and targeted re-asks catch weak fields and unit mistakes. These controls reduce hallucinations and raise trust in outputs for both screening and data extraction stages (Lewis et al., 2020; Wang & Li, 2022; Beurer-Kellner et al., 2024). Surveys on hallucinations list these defenses and show that retrieval and constraints produce strong gains on factual tasks (Huang et al., 2024; Farquhar et al., 2024).

Implementation Patterns That Work

A schema-first design helps. Define fields, types, and units before any prompts. Include examples for edge cases and ambiguous terms. Ask for exact span copies when possible, so reviewers can compare model outputs to the text and settle disputes. Keep prompts short and stable and avoid extra instructions that might distract the model from the schema (Beurer-Kellner et al., 2022).

Retrieval setup comes next. Build a dense index over the corpus. Use proven encoders for text similarity, such as Sentence BERT or SimCSE, and store vectors in FAISS for fast search over millions of passages (Reimers & Gurevych, 2019; Gao et al., 2021; Johnson et al., 2017). At query time, retrieve a few passages for each field and include them in the prompt. Ask the model to cite line numbers or short quotes for each field so a reviewer can check them quickly (Lewis et al., 2020).

Constrained decoding reduces cleanup and failure modes. Use a JSON schema or simples formatting instructions, so the model can only produce valid keys and types and require null values for unknown fields rather than free text (Beurer-Kellner et al., 2024; Park et al., 2025). Self-checking also helps. Run the same prompt several times with slight temperature and sample a consensus. Use a second prompt to verify hard fields, such as sample sizes, units, or confidence intervals (Wang & Li, 2022).

Evaluation should mirror systematic review practice. Use exact match, relaxed match, and per-field precision and recall. Measure agreement with a second human coder to control for ambiguous cases. Flag low confidence fields for review and record the time saved per record for a full cost view. Include an audit trail that links each field to the specific passage used during extraction (Higgins & Green, 2023; Page et al., 2021).

Reporting should follow common norms. PRISMA diagrams can note automated steps, including retrieval and model passes, and the Cochrane Handbook guidance can anchor protocol language about tools and checks (Page et al., 2021; Higgins & Green, 2023). Clear notes help reviewers judge validity and help other groups reproduce the pipeline.

Best Performing Open-Source Models Available Now

Open models have grown strong enough for real extraction work. Mid-size models offer a good balance of speed and accuracy. The Llama family, Mistral and Mixtral, Qwen2 and Qwen2.5, Yi,

and Gemma 2 are common choices with active communities and robust tools (Touvron et al., 2023b; Jiang et al., 2023; Qwen Team, 2024). Public leaderboards help teams compare options for general reasoning and knowledge. The LMSYS Arena and the Hugging Face Open LLM Leaderboard report scores and trends that can guide early model selection (LMSYS, 2025; Hugging Face, 2025).

For extraction tasks, context window length, instruction quality, and output control often matter more than raw benchmark rank. Constrained decoding and retrieval narrow the quality gap between models by anchoring outputs in source text and in a strict schema. A well-designed stack with a strong seven-billion-to fourteen-billion-parameter model can match or beat larger models on structured extraction with lower cost and easier deployment (Beurer-Kellner et al., 2024; Lewis et al., 2020).

Mixture of experts models deserve separate mention. Mixtral variants route tokens through sets of experts, which can offer speed gains when routing is efficient and batching is tuned. These gains can matter when thousands of documents must be processed under a deadline for a grant report or a regulatory review (Shazeer et al., 2017; Fedus et al., 2021).

Why the Llama Family Is a Convenient Choice

The Llama family offers clear model cards, wide quantization options, and many community adapters. That breadth lowers setup time and helps teams find a good fit for local hardware. Meta reports give enough detail to support reproducible runs and fair comparisons across sizes, which matters for academic audits (Touvron et al., 2023a; Touvron et al., 2023b). Newer Llama releases expand context windows, improve multilingual ability, and ship in sizes that fit a range of servers. These traits serve extraction across long documents and mixed language corpora, common in large reviews.

The Llama family also pairs well with efficient fine-tuning. LoRA adapters fit these models cleanly and allow domain tuning on moderate data. QLoRA and GPTQ quantization let teams keep quality while cutting memory and storage, which enables local serving on common workstations (Hu et al., 2022; Dettmers et al., 2023; Frantar et al., 2022). In practice, this means a lab can run extraction over many PDFs using one workstation or a small cluster without sending text to an external provider. That lowers cost and reduces risk for sensitive corpora.

Community support matters as well. Many tools, prompts, and adapters target Llama variants first. This richer ecosystem shortens the path from a test script to a full pipeline that others can repeat. When a team needs to share code and models with partners, shared defaults reduce confusion and speed reviews. Documentation and examples are common, which helps students and new contributors join the work (Touvron et al., 2023b).

Why the Ollama Framework Is a Convenient Implementation

Ollama runs open models locally with a simple command and a small model file. It supports Llama, Mistral, Qwen, and more, and it exposes an OpenAI-style HTTP interface. That interface allows teams to reuse clients and scripts with little change to existing code bases (Ollama, 2025; Ollama, 2024). Local serving protects sensitive text, such as clinical notes, sealed reviews, or partner reports. Logs can be limited to prompts and outputs without raw source text. This setup suits small labs and classrooms where privacy and cost matter.

A typical stack is straightforward. Serve a Llama class model with Ollama, build a FAISS index for retrieval, and decode with a JSON schema or a grammar. Add a small layer for self-checking and consensus, and add a human review tool for flagged fields. This stack runs in a notebook and scales with simple process pools or light orchestrators. Teams can containerize the stack to share it with partners and to ensure consistent runs across machines (Lewis et al., 2020; Johnson et al., 2017).

Ollama’s model files make version control easy. A project can pin a specific build of a model, record its checksum, and keep a copy in local storage. When a new release appears, the team can run a small benchmark on a slice of the corpus and decide whether to upgrade. This practice reduces surprises and supports strong methods sections in papers and reports (Ollama, 2025).

Practical Workflow Example

Consider a review of diet and blood pressure across ten thousand abstracts. The team defines a schema with fields for population, intervention, comparator, outcomes, and study design. The team also defines fields for units, such as millimeters of mercury for systolic change, and sets valid ranges. A short prompt asks the model to fill the fields and copy exact spans for key values, including numbers and units (Reason et al., 2024).

The team then builds a vector index over all abstracts. For each record, the pipeline retrieves sentences that mention diet type, blood pressure values, and follow-up length. The prompt includes those sentences and the schema. The model fills a JSON object and cites the span for each field. Grammar keeps the object valid, and any missing field becomes null rather than free text. A second prompt checks that units match the schema and that effect directions make sense. Records with missing spans or unit mismatches are flagged for human review (Lewis et al., 2020; Beurer-Kellner et al., 2024).

After a pilot on five hundred abstracts, the team measures scores against a gold set. The team tracks exact matches and relaxed match per field, and measures agreement with a second coder. The team also records hours saved per one hundred records and counts errors found in the review step. The pilot results determine whether to scale to all records or adjust prompts, retrieval settings, or the schema. If the run scales, the team version controls all prompts and model settings and stores per-record logs for audits (Higgins & Green, 2023; Page et al., 2021).

A similar approach works beyond health. In materials science, prompts can target composition fields, synthesis steps, and property values. In public policy, prompts can target program names, dates, budgets, and measured outcomes. In education research, prompts can extract sample sizes, grade levels, and assessment types. The same control steps apply across fields: retrieval, schema constraints, self-checks, and human review (Polak & Morgan, 2024; Lewis et al., 2020).

Limits, Risks, and Mitigations

Hallucinations are the most discussed risk. These errors can arise when prompts do not limit the task or when needed passages are not in context. Retrieval reduces this risk by supplying proof text that the model must use. Schema constraints prevent off-schema answers. Self-consistency and targeted re-asks also help catch brittle fields. Reviews survey these tools and show which defenses give the most gain for factual tasks with numeric targets and strict schemas (Huang et al., 2024; Farquhar et al., 2024; Lin et al., 2022).

Privacy presents a second risk. Teams should keep sensitive text local and restrict logs. When remote APIs are used, legal terms should permit processing and storage, and personal data should be masked or removed. Training data leakage is a third risk. Attackers can extract rare strings from models under some settings. Local serving, prompt filters, rate limits, and careful sharing reduce exposure during pilots and production runs (Carlini et al., 2021).

Governance matters for research teams. Model cards and run books should record model version, prompt versions, retrieval settings, and schema definitions. Licensing sets what a team can do with a model and with outputs. The Stanford report on foundation models explains a range of risks and offers guidance for safe use that suits academic labs and industry groups (Bommasani, 2021). For systematic reviews, PRISMA and the Cochrane Handbook offer clear norms for transparent and reproducible steps (Page et al., 2021; Higgins & Green, 2023).

Why LLMs Are a Valid and Effective Method for Unstructured Text

Validity rests on traceable evidence. Retrieval augmented prompts force the model to quote or copy spans from the source. Constrained decoding enforces schema rules and types. Together, these steps cut confabulation and shape outputs into a form that reviewers can check and approve. Multiple results show gains in speed and acceptable accuracy when these steps are used, with the best gains on fields that map to exact spans or simple arithmetic (Lewis et al., 2020; Reason et al., 2024).

Effectiveness rests on coverage and flexibility. Large language models read varied prose and can adapt to new domains with a few examples in the prompt or a small adapter. They also work across languages when weights and tokenizers support that need. When a pipeline must scale across journals, years, and styles, a general model with retrieval and constraints often beats a stack of bespoke rules that demand constant tuning (Minaee et al., 2024; Zhao et al., 2019).

Teams should still measure outcomes with care. Benchmarks for truthfulness and calibration help pick prompts and models for data extraction tasks. Inter-rater metrics show that human review remains essential for certain fields, such as nuanced study design or complex statistical measures that require domain judgment (Lin et al., 2022).

Implementation Notes for Small Labs

A single workstation with a recent GPU can serve a seven to fourteen billion parameter model with a long context window. CPU-only serving is possible with four-bit or eight-bit quantization, at the cost of speed for long prompts (Dettmers et al., 2022; Frantar et al., 2022). Storage needs include the model weights, the FAISS index, and cached runs with metadata for audits. Backups should include prompts, schema files, and small samples of input and output for smoke tests.

Ollama simplifies serving and supports an OpenAI-style API. Python libraries cover retrieval, schema validation, and evaluation. DSPy can compile declarative flows for prompts, retrieval, and checks. LMQL exposes grammar and constraints in a clear language and helps teams keep decoders aligned to a schema (Ollama, 2025; Khattab et al., 2023; Beurer-Kellner et al., 2022). Logging should capture inputs, retrieved passages, and outputs with timestamps and versions so runs can be audited and repeated.

One person can set up a pilot in a week if the schema is ready and the corpus is clean. Larger studies need at least one reviewer per one thousand records for quality control. A run book with clear steps

helps new members join the effort and repeat results across projects. Shared prompts and validation code also reduce errors when many contributors handle different parts of a large review (Higgins & Green, 2023).

Open models and local serving reduce costs. Cloud runs can be reserved for peak loads or long documents. Tracking hours saved and reviewing costs reveals the real benefit, not just raw speed, and supports clear reports. (Page et al., 2021).

Relation to Prior Automated Extraction Tools

Before large language models, tools like RobotReviewer classified abstracts and suggested risk of bias highlights for trials, which cut screening time but did not fully extract structured fields (Marshall et al., 2016). BERT and domain variants like BioBERT improved entity and relation extraction in biomedicine and showed that pretrained language models helped with limited labeled data (Devlin et al., 2019; Lee et al., 2020). Dense passage retrieval replaced sparse methods for many tasks and raised retrieval quality for evidence-oriented prompts (Karpukhin et al., 2020). Sentence-level encoders such as Sentence BERT and SimCSE improved retrieval and reranking quality in many domains (Reimers & Gurevych, 2019; Gao et al., 2021).

These tools laid the ground for current pipelines. Today, teams can combine robust PDF structure parsers, dense retrieval, schema-guided prompts, and constrained decoders to produce repeatable extraction runs. The path from messy text to a clean table now uses fewer moving parts and yields better audit trails than many legacy systems built around rules and custom taggers (Lopez, 2009; Tkaczyk et al., 2015).

Ethical and Legal Considerations

Extraction can mirror bias in the corpus. Teams should test for missing data rates and error rates across subgroups and report gaps in the final dataset. Model selection and prompt design should avoid adding bias by how fields are named and defined. The foundation model report from Stanford lists these risks and suggests plans for safe use across many settings (Bommasani et al., 2021).

Many corpora include licensed or sensitive text. Local serving reduces the spread of such text and helps comply with legal terms and institutional review rules. When public APIs are used, agreements should cover processing and storage, and teams should avoid sending personal data when possible. Research on data extraction from models also shows that rare strings can leak, which supports the case for local serving and careful sharing (Carlini et al., 2021).

Retrieval augmented pipelines make it easier to cite sources for each field. Reports can embed links, quotes, or page numbers so readers can trace each value to the original paper. This practice aligns with PRISMA and similar guidance for transparent evidence reporting and supports reuse of datasets by other groups (Page et al., 2021; Higgins & Green, 2023).

Running very large models carries an energy cost. Open mid-size models and careful batching reduce compute load. Local serving avoids moving large volumes of text to the cloud, which can lower both cost and energy use. These choices also make it easier to meet data residency rules that apply in some regions (Dettmers et al., 2022).

Conclusion

Large language models can extract structured data from unstructured text with strong speed and acceptable accuracy when used with retrieval, schema constraints, and robust checks. Evidence across health and materials science shows that the method works and that it fits existing review norms. Open models now match many needs, and the Llama family offers a practical default due to its tooling and community support. The Ollama framework lowers setup cost and supports local runs that protect privacy. With clear metrics and human review, this approach can support both academic research and applied work in policy, health, and industry.

This chapter offered a map that a small team can follow without deep machine learning expertise. The steps are simple, visible, and testable: define a schema, build retrieval, instruct the model, constrain outputs, and check results. The same plan has worked in related fields such as object detection, where clear pipelines and open tools helped teams move from prototypes to robust systems. In text extraction, the gains lie in speed, scale, and clear audit trails. With measured use and honest limits, large language models can turn messy sources into reliable tables that answer concrete research questions across real research programs.

Object Detection Models and Current State of the Art

Object detection (OD) is a fundamental computer vision task that has become significantly more sophisticated over the past three decades. Early methods employed handcrafted features and conventional pipelines (Haar cascades, HOG [histogram of oriented gradients], and deformable part models) (Edozie et al., 2025). While these techniques based on feature matching were capable of handling variations in illumination, scale, and background clutter, the results frequently exhibited redundant bounding boxes and a high false-positive rate. For example, one of the major advances occurred in 2012 with AlexNet, a deep convolutional neural network (DCNN), which made a significant impact on performance standards (Krichen, 2023). As a result, manually devised features were completely abandoned, and deep learning-based detectors were adopted. These models can enhance the feature representation hierarchy with significantly greater accuracy (Edozie et al., 2025). This revolutionary paradigm became the underlying principle of modern OD paradigms that are powering today's applications in autonomous driving, robotics, medical imaging, and marketing analytics.

Object detection models are theoretically based on the concept of backbone architecture, which is the learning mechanism that focuses on an increasingly abstract hierarchy of feature representation from raw pixels. Region-based Convolutional Neural Network (R-CNN) was the first to propose regions with CNN-based features, but it was computationally expensive (Girshick et al., 2014). Subsequent work improved it and led to Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2015), which utilized region proposal networks (RPNs) in order to speed up the training and inference process for quicker results. On the other hand, one-stage detectors, such as You Only Look Once (YOLO) and Single Shot MultiBox Detector (SSD), achieved end-to-end learning

through direct bounding box prediction and classification, offering advantages in runtime speed (Alkentar et al., 2021). Several recent architectures, such as ResNet (He et al. 2016), EfficientNet, and transformer-based models, have shown better detection accuracy; in particular, the use of transformers has demonstrated very good scalability and generalization ability (Edozie et al. 2025).

The focus on the quality and performance of the final detections has shifted dramatically in recent years, with indices such as Intersection over Union (IoU), mean Average Precision (mAP), and Frame Per Second (FPS) used to evaluate object detection models (Edozie et al., 2025). The recent YOLOv8 version of the models offers competitive mAP performance with a fast inference rate, making it ideal for time-constrained deployments (Ultralytics, 2025). Vision Transformer (ViT)-based models, such as DETR, re-architect the detection pipeline by conceptualizing the detection head as a straightforward set prediction task with custom-designed anchor boxes, which is another innovative concept (Carion et al., 2020). Numerous comparisons have shown that there exists a inescapable trade-off between inference performance and inference speed; the current Yolo designs perform well where the speed requirements of deployment compete with the superiority of accuracy on scale (e.g., DETR-based architectures) (Edozie et al. 2025). The field is thus characterized by a constant negotiation between computational expense, accuracy requirements, and downstream deployment constraints.

Besides technical criteria, object recognition has become an increasingly important subject in scholarly research on both marketing and communication. These models are used by scholars to study consumer behavior, visual attention, and advertising effectiveness by identifying logos,

products, and the human body in images and videos (Edozie et al., 2025). For example, object detection can automatically measure brand exposure at televised sporting events, product placement in movies, and consumer interaction with store displays. OD models have been used in interpreting visual content in social media research and communication studies, where they corroborate large-scale studies of image-based user-generated content (UGC). The growing availability of open-source implementations, especially the YOLO family, has enabled such applications by reducing the barrier to entry for non-technical researchers (Ultralytics, 2025). As a result, object detection has matured beyond its roots in computer vision and has become a methodological tool with numerous transdisciplinary applications.

A very brief literature review will outline the history of object detection models, beginning with handcrafted features and statistical classifiers and continuing through the deep learning revolution that transformed the field in these last decades (Neha et al., 2025). The theoretical basis of research is discussed. The review examines backbone architecture, CNNs *(*convolutional neural networks), residual networks, and models based on transformers. The latter are the technical foundations of contemporary detection systems (Gutierrez et al., 2024). The most important metrics of evaluation are mAP, inference latency, and IoU (Kamal et al., 2024). The section will show how object detection can be applied in marketing studies. The advantage of brand exposure research in the marketing field is the large-scale analysis of content. This is realized with the help of the detection of logos, analysis of consumer behavior, and the analysis of communication strategy (Hosseini et al., 2025). The key takeaway from the review is the trade-off between speed, accuracy, and usability in YOLO, which has contributed to its popularity in both applied industry use within the community and academic research (Ali and Zhang, 2024).

History of Object Detection Models

The history of object detectors dates back to classical computer vision, which was, to a considerable extent, based on manually crafted features and other classical machine learning classification techniques. Early feature extraction methods, including Scale-Invariant Feature Transform (SIFT), Canny edge detection, and Histogram of Oriented Gradients (HOG), helped to solve the problem of objects with varying brightness, size, and orientation (Nguyen et al., 2014; Neha et al., 2022). Engineering algorithms were used to extract local features into fixed-length vectors, which were then classified by models such as support vector machines (SVMs). Although they formed a basis for deriving higher semantic understanding, they only performed sufficiently well in controlled environments and not in cluttered or generalized settings (Neha et al., 2022). These handcrafted pipelines, which blended low-level image features with contextual heuristics, lacked the representational capacity required for robust and generalizable detection (Neha et al., 2022).

One key advance in this initial period was a real-time object detector, the Viola-Jones face detector, which employed Haar-like feature detectors and a cascade of boosted classifiers to detect objects in a video stream at high speed (Taunk et al., 2020). This breakthrough demonstrated that detection could be performed at interactive speeds, setting the stage for its use in surveillance and human-computer interaction. However, the technique was not well generalized, as it worked best with rigid and highly structured targets such as frontal faces. Later methods, such as the Histogram of Oriented Gradient (HOG) + SVM detector, were extended to pedestrian detection, while the Deformable Part Model (DPM) utilized part-based features to capture intra-class variance (Galvao et al., 2021). These models were computationally expensive, designed using manual feature engineering, and suffered from inefficient sliding-window searches (Neha et al., 2022). These

limitations represented the importance of a more powerful data-based feature extractor that can implicitly learn hierarchical image representations.

In 2012, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) heralded the deep learning revolution in computer vision when a model known as AlexNet (Krizhevsky, Sutskever, and Hinton, 2017) delivered exceptional performance in the challenge. The advancement proved that convolutional neural networks (CNNs) were significantly better at image classification then previous models, and soon after, object detection models started switching to CNN-based backbones (Girshick et al., 2014). R-CNN first introduced this novelty, achieving significant improvements over handcrafted pipelines by combining region proposals with CNN-based feature extraction. Although performing, R-CNN was computationally costly due to the use of selective search and independent feature extraction per proposal. Over time, models such as SPP-Net (He et al., 2015) and Fast R-CNN (Girshick, 2015) refined this process, sharing convolutional feature maps between proposals and therefore allowing for end-to-end training and running more efficiently. The idea to propose objects using shared feature maps (Region Proposal Networks (RPNs) of Faster R-CNN) (Ren et al., 2015) was a move in the right direction of efficient and high-quality object detection. These two-stage CNN-based approaches, as mentioned by Neha et al. (2022), overcame numerous drawbacks of classical approaches by training hierarchical, semantic-abundant features that can handle complex scenes with occlusion and variable object scale.

In parallel to these models, one-stage detectors that favor real-time performance without the use of explicit region proposals were proposed. The You Only Look Once (YOLO) model predicted

bounding boxes and classes directly from grid cells (Redmon et al., 2016). The processing speed exceeded 45 frames per second (fps), allowing the model to be used for online inference processing. This concept was later expanded to the Single Shot MultiBox Detector (SSD) (Liu et al., 2016). SSD adapted multi-scale feature maps to different-sized objects. However, it was also slightly less accurate, but while maintaining high speed. According to Neha et al. (2025), this marked a paradigm shift from feature engineering to representation learning. This change provided the basis for the modern transfer-based networks as well as real-time systems that are currently leading object detection research.

Backbone Architectures in Object Detection

Object detection models are built based on backbones. They are used as feature extractors, which convert raw image information into a format that can be used in detection tasks. This space is dominated by CNNs, which process spatial hierarchies of features, starting with edges and progressing to more complex patterns (Bouraya & Belangour, 2021). The backbone chosen has a direct influence on the detection model accuracy, speed, and generalization capabilities. Research showed that substituting a conventional backbone with a heavier or deeper one tends to lead to significant improvements in accuracy but also higher computational cost (Bouraya & Belangour, 2021). Nowadays, this constitutes a truism for backbones in both experimental studies and applied object detection.

Previous object detection methods were based on handcrafted features and the use of statistical classifiers. Local descriptors were then obtained through the application of various techniques such as HOG and SIFT, and eventually implemented in conjunction with sliding windows and SVMs

(Kalake et al., 2022). These techniques worked at the time, but were sluggish and unable to handle more complicated visual variation. Deep learning has been a breakthrough in this area, as it introduced feature extraction within the network (Sumit et al., 2024). Fast R-CNN proved to be more efficient compared to Mask R-CNN, where the proposal stage was added to the backbone (He et al., 2022). Mask R-CNN was developed based on this pipeline, incorporating segmentation, further showcasing the effectiveness of CNN-based backbones on different tasks (Bouraya & Belangour, 2021).

The considerable number of architectures developed proved the ductility and adaptability of backbone designs. The high accuracy on ImageNet dataset of AlexNet initiated the deep learning wave in 2012 (Bouraya & Belangour, 2021). The depth of VGG16 was optimized through the use of uniform convolutional layers at the cost of computation. The use of residual connections to design networks was truly revolutionary and led to the possibility of training networks with hundreds of layers. GoogleNet introduced inception modules, consisting in features extraction at varying scales (Boesch, 2024). Based on the same idea, in DenseNet each layer was connected to all the preceding ones to improve gradient flow and promote feature reutilization. MobileNet exploited depth wise separable convolution to ensure that the network remained lightweighted, a design choice that enabled it to be deployed on mobile and embedded platforms.

The trade-offs between precision, speed and depth can be identified best through comparative studies. MobileNet and ResNet18 are fast and inexpensive to train models that deliver good results for tasks like surveillance or marketing applications (Shahriar, 2025). In comparison, bigger

networks such as NasNetLarge and SeNet154 are more accurate and take more time to be trained, with significantly higher computational costs (Bouraya and Belangour, 2021). ResNeXt and InceptionResNetV2 intermediate designs are better in terms of average efficiency. Deeper networks tend to enhance accuracy at the cost of inference latency. Applied fields are particularly sensitive to this trade-off, with large-scale logo recognition or consumer engagement experiments requiring models that offer reasonable accuracy and can be deployed as fast as possible.

The systematic review implemented by Aziz et al. (2020), which analyzed more than 300 papers in the field of deep-learning-based object recognition, looked at two general groups of detection systems: region proposal-based systems, like R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, and regression or classification-based systems, such as YOLO, SSD, RetinaNet, and EfficientDet. Aziz et al. (2020) documented the trade-off between accuracy, computational cost, and real-time performance by mapping these architectures over performance indicators and core characteristics of the models, including bounding-box regression, feature extraction strategies, and network depths.

One of the main contributions from Aziz et al. (2020) was identifying technical limitations that are still present in detection pipelines, such as intra-class variation, scale sensitivity, and the trade-off between the inference latency and detection performance. Even if associated with high computational cost, R-CNN architectures showed high accuracy improvements through the use of region proposals. Single-shot models like YOLO and SSD were fast, but had difficulties with very small targets unless carefully tuned (Arwidiyarti, 2025). Comparisons of this sort emphasize the

convenience of standard benchmark measures as mean Average Precision (mAP) and Intersection over Union (IoU). These are the metrics of preference to determine quality in detection across datasets.

Aziz et al. (2020) describe object detection applications in surveillance, medical imaging, transportation, and consumer tools, showing that the tasks these models are applied to have grown past general detection. More specialized properties like face recognition and salient object detection, as well as pedestrian tracking were tackled. This convinced researchers of the validity and appropriateness of the field for marketing and communications research, where the detection of brand logos, visual interaction analysis, quantification of the consumers' interactions required a quantifiable and efficient method of inquiry (Aziz et al., 2020).

In recent times a growing demand for lightweight object detection models has emerged. Models are increasingly targeted for implementation on edge devices, where memory, power, and latency are major constraints (Mittal, 2024b). Lightweight detectors make the most of their efficiency and accuracy to allow them to be deployed on IoT platforms, mobile devices, and embedded processors. The computation partition, redundant-layer pruning, and neural architecture search can reduce the size of the network without significantly reducing its accuracy (Mittal, 2024b). This trend suggests that more lightweight architectures will obtain competitive performance on benchmarks, such as COCO and Pascal-VOC, and be usable in real-time work in healthcare, retail, and autonomous systems.

Backbone effectiveness is also defined by the training strategy. In some models, pretrained models can be object of further refinement. In this case, batch normalization is adopted to stabilize training with the support of other models (Zhu et al., 2019). This makes models less dependent on popular datasets like ImageNet, but adds a more time-consuming training step. Other post-training enhancements are possible. DetNet is an example in which kernels are dilated to create larger receptive fields, without attempting to downsample feature maps (Li et al., 2018). While improving the performance of the models, these methods are more resource-intensive and are not always preferred to pretrained models, except in cases that require customization.

Head network improvement focuses on feature learning that is independent from the main task, while keeping the backbone still relevant. The subnetworks of classification and localization used in the RetinaNet illustrate this dynamic where accuracy plummets in the absence of these subnetworks (Zhang et al., 2024). Subsequent models, such as Double-Head RCNN and Task-aware Spatial Disentanglement (TSD), improve on this premise by leveraging features devoted to specific subtasks (Zhang et al., 2024; Wang & Li, 2022). Separating features on the basis of subtasks and avoiding mixing classification and localization showed to be an effective strategy (Song, Liu & Wang, 2020).

Backbone optimization can be conducted in three prevalent ways: network redesign, training strategy, and head refining, but in the end all methods have to balance between precision, performance, and the cost of training (Zhang et al. 2024).

Table 1 *Thematic Summary of Backbone Architectures in Object Detection*

Theme	Key Contributions / Models	Strengths	Limitations / Trade-offs	References
Foundations of Backbones	• CNN-based feature extractors • Spatial hierarchies: edges → complex patterns	• Strong accuracy • Robust feature extraction	• Heavier models ↑ computation • Complexity grows with depth	Bouraya & Belangour (2021)
Early Approaches (Pre-CNN)	• Handcrafted features (HOG, SIFT) • Sliding windows + SVMs	• Useful for simpler tasks • Established groundwork	• Slow & rigid • Poor at complex variations	Kalake et al. (2022); Sumit et al. (2024)
R-CNN Family	• R-CNN, Fast R-CNN, Mask R-CNN	• Region proposals improve accuracy • Segmentation via Mask R-CNN	• High computational cost • Latency in inference	He et al. (2022); Bouraya & Belangour (2021)
Key Architectures	• AlexNet → sparked deep learning • VGG16 → uniform conv layers • ResNet → residual connections	• Accuracy ↑ with depth • Feature sharing enhances learning • Mobile deployment feasible	• Computation-heavy (VGG, ResNet) • Lightweight → lower accuracy ceiling	Boesch (2024)

	• GoogleNet → inception modules • DenseNet → feature sharing • MobileNet → lightweight, mobile			
Comparative Trade-offs	• MobileNet, ResNet18 → speed • SeNet154, NasNetLarge → accuracy • ResNeXt, InceptionResNetV2 → balance	• Lightweight → real-time apps • Deep models → state-of-art accuracy	• Lightweight → less accurate • Deeper nets → latency & power issues	Shahriar (2025); Bouraya & Belangour (2021)
Survey Findings	• Region-based: R-CNN, Fast/Faster/Mask R-CNN • Regression-based: YOLO, SSD, RetinaNet, EfficientDet	• High accuracy (region-based) • Real-time speed (YOLO/SSD)	• Region-based: slow, costly • Single-shot: weak at small objects	Aziz et al. (2020); Arwidiyarti (2025)

Technical Challenges	• Intra-class variation • Scale sensitivity • Trade-off: latency vs accuracy	• mAP & IoU serve as strong benchmarks	• High cost • Difficult tuning for small-scale	Aziz et al. (2020)
Applications	• Surveillance • Transport • Health • Consumer/retail (logos, engagement)	• Strong real-world impact • Specialised detection tasks	• Deployment depends on a balance between speed & accuracy	Aziz et al. (2020)
Lightweight & Future Trends	• NAS-FPN, DCNv2 • Neural Architecture Search • Layer pruning & partitioning • IoT/mobile deployment	• Efficiency for edge devices • Competitive performance on COCO, Pascal-VOC	• Smaller capacity • Still catching up with deeper nets	Mittal (2024b); Aziz et al. (2020)
Backbone Optimization (Figure 1)	• Feature Pyramid Network (FPN) • Path Aggregation Network (PANet) • STDN, M2Det • DetNet, ScratchDet	• Multi-scale feature fusion • Preserves spatial info • Independent task branches' ↑ accuracy	• Pretraining removal = resource-heavy • Some designs are complex to implement	He et al. (2015); Zhou et al. (2018); Zhao et al. (2019); Li et al. (2018)

	• Double-Head RCNN, TSD			

State of the Art and Current Best Models by Relevant Metrics

Benchmarks and performance metrics

Common examples of standardized metrics that are used to measure the state of the art in object detection are Intersection over Union (IoU) and mean Average Precision (mAP) (Chahal & Dey, 2018; Henderson & Ferrari, 2016; Ali & Zhang, 2024). These are the same indicators that are being used in competitions on popular benchmarks such as COCO and PASCAL VOC. The framework centered on competition and open-access challenges sparked a decade of discovery in detection research projects (Everingham et al., 2010; Xin et al., 2024). IoU is a simple metric used to determine the location accuracy of the bounding boxes predicted in relation to the ground truth (Putra et al., 2025; Wang & Wu, 2021). It measures the degree of overlap between the predicted bounding box and the ground truth on the annotated image. Thresholds of $IoU \geq 0.50$ or $IoU \geq 0.75$ have been defined as standard cutoffs for accurate object localization depending on the specific task (Padilla et al., 2021). On the other hand, mAP measures the aggregate performance of models across different object classes, by measuring the area under the precision-recall curve for a single class and a specific Intersection over Union (IoU) threshold, constituting the preferred metrics for inter-model comparisons.

Recent sources point to the fact that mAP, in general, and its COCO-style version (mAP@[.5:.95] specifically) is considered the gold standard for measuring object detectors. It does not reward

models that perform well at lower thresholds only, but also those that do not achieve fine-grained localization (Zou et al., 2023), since detectors with good accuracy at IoU = 0.5 might not maintain the same level at IoU = 0.75 and above. Current studies focus on those models that perform well on a series of different thresholds, which guarantees good generalization in practice (Wu, Li & Wang, 2020). The application of these metrics has been crucial in the automotive driving industry as well as in medicine since these are fields that require generalizability and flexibility.

Popular Models

Faster R-CNN, the YOLO series, and RetinaNet are the most popular and frequently used models nowadays, as shown by Tan et al. (2020). In Faster R-CNN, a region proposal mechanism was introduced with a major impact on the detection performance, especially in a high-IoU environment, where localization accuracy is highly important (Ren et al., 2015). The YOLO framework focusses on real-time detection and implements single-stage prediction techniques (Ali & Zhang, 2024; Alhassan & Yilmaz, 2025). Bochkovskiy et al. (2020) illustrated that these models perform well where inference speed is paramount without considerable losses in mAP, when tested on the COCO benchmark. On the other hand, RetinaNet was able to overcome the class imbalance issue with the help of focal loss, which enhanced object detection for smaller objects and increased mean precision on complex datasets (Lin & Chen, 2024).

EfficientDet constitutes another step forward and can be considered a breakthrough regarding the accuracy-efficiency trade-offs. The model uses a compound scale technique to scale depth, width, and resolution independently, allowing it to be used effectively at different scales and under tight

computational constraints. EfficientDet achieved competitive scores for mAP@[.5:.95], while the size of the detector was kept small enough to be deployed on edge devices (Tan et al., 2020). In accordance with the raising importance of computational efficiency in the field, Padilla et al. (2021) proposed listing measuring detection speed and robustness, along with accuracy, as relevant evaluation metrics.

Transformers

Transformers such as DETR have introduced attention-based models that have achieved competitive mean precision results, paving the way for a significantly shorter detection pipeline (Carion et al., 2020). DETR showed that by removing heuristic aspects of the algorithm, such as non-maximum suppression, it was possible to still obtain state-of-the-art results, but the considerable training cost and inefficiency remain a disadvantage (Gao et al., 2021). In its latter versions, such as Deformable DETR, this issue was mitigated but not resolved (Zhu et al., 2019). These new architectures point toward interpretability and scalability optimization, without renouncing to competitive IoU and mAP results (He et al., 2025).

Current best models are not identified by one metric alone, but evaluated using a collection of benchmarks that use a combination of the IoU thresholds, mAP across scales, recall and latency in some cases. Mekhalfi et al. (2021) show that YOLOv5 and EfficientDet are superior in applications that require speed and resource-efficient execution, whereas Faster R-CNN, RetinaNet, and DETR are found to be the better choices with more stringent IoU conditions (Sapkota et al., 2024). Since the evaluation metrics and the evaluation formats may vary, as Padilla et al. (2021) noticed,

necessity to use common benchmarking tools arose. The open-source toolkits offered by projects such as COCO and Pascal VOC can offer a crucial framework that enables researchers to test detectors in a consistent manner, allowing them to cross-check studies and results more effectively and transparently. In this environment the enterprise for a state-of-the-art model can be thought of as moving across a continuum of metrics in which the goals for faster, most precise, and generalizable models are tackled as one task.

Table 1 *Thematic Summary of State-of-the-Art Object Detection Models & Metrics*

Theme	**Key Elements / Models**	**Strengths**	**Limitations / Trade-offs**	**References**
Evaluation Metrics	• IoU (≥0.50, ≥0.75) • mAP (COCO-style [.5:.95])	• IoU → localization accuracy • mAP → class + threshold generalization • Gold standard benchmarks	• IoU alone is insufficient • Some models drop at higher IoU thresholds	Padilla et al. (2021); Zou et al. (2023)
Popular Models	• Faster R-CNN • YOLOv4, YOLOv5 • RetinaNet	• Faster R-CNN → high IoU precision • YOLO → real-time speed	• Faster R-CNN → slower inference • YOLO → small-object weakness	Ren et al. (2015); Bochkovskiy et al. (2020); Lin & Chen (2024)

		• RetinaNet → focal loss for imbalance	• RetinaNet → heavy compute cost	
Efficiency-Oriented Models	• EfficientDet	• Compound scaling (depth, width, resolution) • Competitive mAP@[.5:.95] • Lightweight, deployable	• Limited capacity on very large datasets	Tan et al. (2020); Padilla et al. (2021)
Transformer-Based Models	• DETR • Deformable DETR	• Attention mechanism • Removes heuristics (e.g., NMS) • Competitive mAP	• DETR → slow convergence • Deformable DETR → resource demands still high	Carion et al. (2020); Zhu et al. (2019)
Benchmarking Practice	• COCO • PASCAL VOC • Padilla Toolkit	• Transparent comparison • Multi-threshold testing • Ensures reproducibility	• Benchmarks may miss domain-specific context	Padilla et al. (2021)

Small-Object Detection	• Context-aware pipelines • Spatial/semantic priors	• Better recall of tiny objects • Reduces false negatives	• Standard mAP fails to capture fine-grained gains	Martinez-Ríos et al. (2022); Mi et al. (2025)
Cross-Domain Robustness	• Scene-sensitive models • Semantic generalization	• Works across domains (street → aerial) • Resilient to scale, background change	• Needs stronger adaptation frameworks	Jamali et al. (2025)
Temporal Consistency	• Frame-level accuracy • Temporal coherence metrics	• Stable predictions in videos • Preserves object identity • Reduces flicker/jitter	• Ignored in IoU/mAP-only metrics	-
Industry Applications	• Autonomous driving • Medical imaging • Marketing analytics (logo detection, engagement)	• High contextual reliability • Task-specific adaptation	• Context-agnostic metrics → overestimate reliability	-

Future Trends	• Hybrid evaluation • Combine IoU, mAP, speed, context scores	• More comprehensive view • Aligns with real-world needs	• Increased complexity in benchmarking	Padilla et al. (2021); Jamali et al. (2025)

More recently, object detection has advanced to the point of being a distinct field, not relying on a single measure to evaluate model performance but considering Intersection over Union (IoU), mean Average Precision (mAP), and inference latency. The metrics that are used for comparing models are still traditional ones, including IoU and mAP. Nonetheless, they fail to capture other problems such as small object filtering, domain shifts, and velocity consistency in video tasks. Models such as Faster R-CNN, YOLO, RetinaNet, and EfficientDet have various trade-offs among speed, accuracy, and efficiency. Individual optimization methods can rebalance these characteristics but also introduce significant bias (Lin & Chen, 2024; Tan et al., 2021). Models such as DETR that involve transformers are signs of a change in development direction (Feng et al., 2023) toward models where the complexity of the pipeline is reduced by attention mechanisms without sacrificing accuracy (Aromoye, Hiung & Sebastian, 2025; Yu, Tang & Mu, 2025). Based on these comparisons, it is clear that no model will be most effective in all cases. Model choice goes along with the application, computational capability, and the scale of implementation.

These insights are backed by use cases in industry and research. Self-driving systems require temporal consistency and a developed understanding of the scene. Medical imaging depends on the ability to pick out small or subtle features (Hussain et al., 2022; Pinto-Coelho, 2023).

In the marketing space, big data on consumer impressions and visual communications can be studied through accurate detection of brand logos and brand elements under conditions of media saturation (Hosseini et al., 2025). The trend in current research is that mixed-evaluation models, combining IoU and mAP with context-aware and temporal performance scores, will become standard (Ghanaei & Rouhani, 2025). This synthesis indicates that the field is leaving general precision numbers aside and focusing on models with the ability to process real-life scenarios, as well as measures that ensure technical gains and ecological validity in the identification of scenarios by object detection systems.

Computer Vision Models Applications in Marketing Studies

Computer vision (CV) models are becoming methodological tools for advertising content analysis within marketing and communication sciences (Haleem et al., 2022; Lyndyuk et al., 2024). Manual coding can produce rich sociocultural data; however, it is limited in scalability and can also cause analyst burnout (Li and Zhang, 2024). By contrast, CV models enable researchers to work with high volumes of images and isolate patterns and associations otherwise difficult to perceive through manual analysis. This marks a shift in the methodological frontier of advertising studies. Working with these methods at scale and applying them to study the implications of visual content for consumer involvement and brand expression becomes efficient, consistent, and repeatable.

New specialized models can tackle completely new tasks like emotion recognition, which shifts advertisement analysis from a content identification rubric to the consumer's affect. It can be performed through DeepFace or FaceReader, automatic facial expression recognition and classification tools. These instruments help investigate numerous emotional responses elicited by

the brand message (Li and Zhang, 2024). The results offered by the models are measurable (e.g., discrete emotion probabilities) and can be adjusted to fit survey or behavioral data. This enables triangulation, whereby visual evidence of emotion can be used to disclose or challenge self-reported attitudes. Emotion recognition has been considered in advertising research. It has made advertising research more detailed, in the sense that methodological instruments can answer questions about the impact of affective cues on liking, trust, and purchase intentions toward brands.

Customer movement in the retail sector may be reviewed with the help of object detection. Retailers generate a massive volume of CCTV footage daily, but much of the footage is wasted. It can be fed into computer vision, which enables businesses to transform raw video into quantifiable information. This is not passive retail monitoring but an active approach to retailing that aims to generate insights. Through OD models implementation meaningful variables such as the number of people passing by a place, the time they spend at the place, and the conversion rate can be quantified (Javare et al., 2020). These measures are used to track the movement of customers in a store. Aggregated detection data reveal high-traffic and underused regions. Managers are then able to modify layouts and the positioning of products and advertising displays depending on the observed movement behaviors.

Object detection presents a tremendous use case for facilitating data-driven marketing research (Khan & Imran, 2024; Sun, Sun & Chen, 2024). It offers scalability, replicability, and objectivity, which too frequently does not apply to manual observation. The interaction between visual analytics and statistical techniques is an example of a hybrid method wherein the derived

qualitative information is supported by quantitative data derived by detection models (Guetterman, Fetters & Creswell, 2015; Noyes et al., 2019). CV models are enablers for new methodological approaches in advertising and communication studies. They allow researchers to switch to large-scale computation. Current models are not concerned with universal fitting; rather, researchers have consciously chosen to adjust their study models (Li & Zhang, 2024). This means critical reflexivity: an understanding that research results are also a product of modeling choices, and that some quantification methods, such as IoU and mAP, have to be thought of as salient in the landscape of marketing inquiry. One way of understanding the use of these methods is that CV models are not merely instruments, but a way of doing things that can inform academic research on marketing and communication with new significance.

Focus on YOLO Models

The first YOLO model in 2015 changed the way object detection can be performed by considering the detection process as a regression problem (Cong et al., 2023). Unlike two-stage detection, such as the Faster R-CNN, YOLO predicted bounding boxes and classes in one go (Mohammed, 2025). The approach overcame issues of speed and made use in real time possible (Murat and Kiran, 2025). Later versions would refine the design. YOLOv2 introduced anchor boxes while YOLOv3 introduced the network of residual and feature pyramid networks (Kang et al., 2025; Pebrianto et al., 2023). These changes resulted in slight accuracy degradation but still kept YOLO faster than many of its rivals in domains like autonomous driving and medical imaging (Ali & Zhang, 2024).

A major strength of YOLO is its flexibility. The model is extended far into domains where more data requirements and constraints are to be considered. Researchers have successfully applied YOLOv4 and YOLOv5 in medical tasks, such as identifying tumors, skin lesions, or fractures, and have obtained high-level accuracy with minimal processing costs (Murat & Kiran, 2025). Likewise, YOLO has had a major role in the agricultural automation systems, with adaptations of algorithms, such as YOLOv7, used to detect fruits, estimate their ripeness, and control robotic harvesters (Shaikh et al., 2024). These examples demonstrate that even with use cases that require both precision and computational efficiency, single-stage detectors like YOLO continue to be useful.

Yet criticalities concerning YOLO models in specific tasks, such as small object detection and complex scene generalization represented a challenge (Murat & Kiran, 2025). Although YOLOv1 and YOLOv2 were not performing well in cluttered settings, recent versions like YOLOv8 and YOLOv11 have added new mechanisms like oriented bounding boxes (OBB), GELAN, and probabilistic generative inferences (PGI) (Jegham et al., 2024). These developments broke the negative feedback between advancements in model architecture and efficiency. Overall, they point toward improving computer vision where models are becoming increasingly idealistic regarding hardware choices as well as task-specific challenges (Bochkovskiy et al., 2020).

A second general theme in recent papers is to compare YOLO with other real-time object detectors. Examples include SSD and RetinaNet, which offer competitive baselines in accuracy. Nevertheless, they are not always faster than YOLO in FPS (frames per second), and this characteristic makes YOLO more likely to run in environments where processing time is a sensitive

parameter, such as surveillance or robotics (Wang & Li, 2022). The speed–accuracy trade-off is a persistent issue, but the philosophy behind YOLO's design has emphasized meeting speed requirements for embedded equipment (Wei et al., 2025). This design choice makes it common in embedded systems, drone applications, and IoT applications.

Measures such as mAP and IoU are still important measures in the evaluation of YOLO's models in terms of performance evaluation (Ayachi et al., 2025). Progress in FLOPs efficiency, particularly for YOLOv9 to YOLOv11, has reinforced the case for their adoptions in areas where computing resources are not abundant (Murat & Kiran, 2025). For example, YOLOv11 has better compatibility with less memory based on GPUs and good performance on a variety of datasets (Kishor, 2024). Such a balance can be credited to the maturity of the model, as not only is it an innovative algorithm, but it also considers the practicality of deployment.

Lastly, the continuous evolution of the YOLO framework represents a broader trajectory within the field of deep learning. Its applications extend across multiple domains, including biology, behavioral analysis, and various areas requiring adaptability and domain-specific optimization. A comparative overview of the YOLO versions enables a structured assessment of their respective contributions beyond technical innovation alone. Recent iterations, such as YOLOv11, have been developed to address the limitations of earlier models, indicating a gradual consolidation of research that positions YOLO as a central paradigm in contemporary object detection studies (Ali & Zhang, 2024). Owing to its speed, accuracy, cost effectiveness, YOLO continues to maintain a prominent role among the preferred frameworks adopted in scientific, industrial, and applied research contexts.

Section Summary

This section outlined the evolution of object detection from handcrafted methods such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and Deformable Part Models (DPM) to contemporary deep learning-based architectures that now define the state of the art. Classical methods laid the groundwork for feature extraction and object recognition but were limited by weak generalization, high computational demands, and reduced performance in complex environments. The introduction of Convolutional Neural Networks (CNNs), beginning with architectures such as AlexNet and R-CNN, represented a major turning point. Two-stage detectors, including Faster R-CNN, achieved high levels of accuracy but at a considerable computational cost. Real-time detection frameworks such as YOLO and SSD subsequently emerged to balance precision and inference speed, enabling broader practical deployment.

Backbone architectures, including ResNet, DenseNet, MobileNet, and EfficientNet, have been central for improving performance by optimizing depth, connectivity, and computational efficiency. Complementary advances such as Feature Pyramid Networks (FPN) and transformer-based architectures, notably DETR, have shown enhanced detection performance across object scales and domains. Despite these architectural variations, Intersection over Union (IoU) and mean Average Precision (mAP) remain the standard evaluation metrics within the field.

Applications across autonomous driving, healthcare, agriculture and marketing illustrate the transversal potential of these models. Case studies in industry show that YOLO architecture

demonstrates how incremental improvements can address real-world challenges, shifting emphasis from purely speed-optimized models to architectures that integrate efficiency with robustness and precision. This was evident also from how, within communication and marketing research, computer vision models such as YOLO have extended beyond technical functions to serve as analytical tools capable of examining consumer behavior at scale, facilitating systematic analysis of brand visibility, audience engagement, and the dynamics of visual participation in a consistent and replicable manner.

In this work YOLOv11 will be our model of choice. Since objects in social media posts are rarely hidden in cluttered scenarios, especially if the influencers use them as communicational props, YOLOv11 can operate under optimal conditions, while keeping computational cost low and allowing for the use of local devices to precede with the analysis.

Thematic Analysis of Finfluencers' Text in Social Media Posts

The introduction of social media has transformed the way information, specifically in personal finance, is disseminated. With its visually appealing structure and extensive user base, Instagram has become a leading platform used by financial influencers, or finfluencers, who present themselves as democratizers of financial advice (Mölders et al., 2024). Such content creators capitalize on the use of short-form content, including reels and stories, to simplify complex financial concepts and reach a large audience, typically young demographics such as Gen Z and millennials, who have historically had limited access to professional financial advice (Wahyudi et al., 2025). The elevation of finfluencers on Instagram has been attributed to the algorithm that encourages relatable and bite-sized content on the platform, enabling creators to reach a large audience and influence financial behavior at scale (Boston Institute of Analytics, 2025). For example, finfluencers frequently post stories about financial empowerment, encouraging their followers to invest, save, and work toward financial independence through methods such as stock trading or cryptocurrency business ventures, thereby reducing the divide between traditional financial institutions and their users (Moreish, 2025).

However, this democratization is accompanied by increasing concern about misinformation and regulatory loopholes. Often, finfluencers are unqualified, and therefore, they propagate false or biased information that could misdirect new investors. According to regulatory bodies, there have been cases of manipulative activities by social media influencers who have promoted stocks without disclosing their ownership ties and have been involved in schemes that negatively affected retail investors (Hitchcock, 2023). Issues related to these have been explicitly mentioned in the 2023 enforcement results of the U.S. Securities and Exchange Commission (SEC), which reported

instances of social media finance influencers involved in securities fraud cases (Mittal, 2024a). Moreover, rules, laws, and regulations are not keeping pace with the rapid growth of online platforms, leaving loopholes that allow influencers to offer unregistered investment recommendations, thereby increasing the risk of pump-and-dump schemes and data theft. A gap exists in recent and large-scale studies of finfluencer content on Instagram (Hasanah et al., 2025). The literature typically focuses on small groups or alternative applications, including TikTok, which creates knowledge gaps in the data-driven understanding of Instagram-related narratives and their implications for financial literacy.

Research Aim and Questions

An exploratory methodology is employed in this research to analyze the content published by finfluencers on Instagram, identifying the primary themes and narratives that may influence the audience's perception of finances. This section is primarily aimed at examining the latest trends and stories in Instagram posts about finfluencers, as these provide a subtle insight into how these influencers discuss the financial aspects of a digital environment.

The following research questions are suggested to direct this study:

1. What do finfluencer captions on Instagram discuss most of all?
2. How do these themes manifest as larger financial empowerment discourses?
3. Does a large language model (LLM) serve well to support exhaustive thematic analysis (TA) of large datasets?

These questions revolve around influencer post content and the methodological promise of artificial intelligence in qualitative studies, particularly in the context of content analysis and technological innovation.

Novelty and Contributions

The research is timely, as it is based on data from the current Instagram ecosystem in 2024 and 2025, which indicates the evolving trends in a dynamic online environment. The depth of the analysis is respectable, based on a dataset of 22.854 posts accessed using a custom web crawler, hereafter referred to as the *Instagram Finfluencers Database*. This large corpus enables a robust study of patterns that would be unfruitful with smaller datasets. The research methodology is innovative in that it employs an LLM-based thematic analysis, which replicates human coding processes to efficiently process massive data volumes while still allowing for detailed coding. This not only makes it more scalable, but it also demonstrates the usefulness of AI in qualitative research, breaking the limitations of traditional manual TA with big data.

Its contributions are threefold: empirically, it provides novel insights into the topics discussed on Twitter by Instagram influencers; theoretically, it relates these topics to empowerment discourse in social media finance; and practically, it justifies the viability of LLM in providing TA at scale, offering a prospective replicative design for future researchers.

The section is organized as follows: following this introduction, the literature review presents the published research on the issue of finfluencers, misinformation, and financial narratives, justifying the use of the LLM to conduct the analysis. The methodology section outlines the data gathering

process, including the use of the custom scraper, the TA procedure assisted by the LLM, and ethics considerations. Results indicate the themes identified, their prevalence, and correspondence to the discourses of empowerment. The discussion supports the results with previous literature, examines the implications of regulation, and assesses the effectiveness of LLM. Lastly, it concludes by summarizing important findings, limitations, and future research directions.

Thematic Analysis in Social Media Research

Thematic analysis (TA) is one of the most widely recognized approaches to qualitative research, characterized by its versatility in identifying, analyzing, and presenting patterns in data (Ahmed et al., 2025). In 2006, Braun and Clarke initially defined TA as the provision of a systematic yet adaptable framework that allows researchers to operate in a qualitative context without necessarily adhering to specific theoretical presuppositions (Braun & Clarke, 2006). This is a six-step process that involves collecting information about the data, pre-coding, identifying themes, reviewing themes, defining and naming themes, and producing a report (Braun & Clarke, 2006). The method is also beneficial because it is easy to use, allowing both less experienced researchers and experienced researchers to find it productive (Braun & Clarke, 2022). They later revised it by simplifying the methodology and making TA a reflexive process, where the subjectivity of the researcher is viewed as an asset rather than a prejudice that should be minimized (Braun & Clarke, 2022). The development of such a form accentuates its adaptability to different positions in epistemology, including realist and constructionist, making TA more feasible within a wider variety of fields.

TA is effective in the context of social media research, particularly in the study of the vast and diverse content published on social media networks such as Instagram and Twitter (De Castro, 2023). TA is reliable in recognizing patterns of social media data, which are characterized by conciseness, multimodality, and user-generated content (Zachlod et al., 2022). In particular, the application of TA to analyze the discourse of health-related space influencers has been utilized by scholars keen on discussing how influencers can impact the population's opinion, sharing their personal accounts and advice (Michel et al., 2024). The influencers have been transformed into professionals in the unregulated fields (Flaherty & Mangan, 2025). TA has been applied to themes in the teaching tools of financial literacy, identifying patterns of framing financial advice to empower or educate the audience (Bhatia et al., 2024). The fact that most studies on financial literacy, risk management, and behavioral economics share common trends suggests that these trends are recurring, but can be attributed to a significant extent to assumptions concerning demographic factors (Khan et al., 2022). These applications demonstrate that TA can be successfully employed to the extent that the size and character of the information stored in social media posts offer informative findings on emerging stories that inform user behavior.

Exploratory TA is particularly appropriate when the purpose is producing initial insights from extensive datasets where pre-existing categories may not be sufficiently applicable (Wittmann, 2024). Unlike deductive research approaches, where hypotheses are tested, exploratory TA allows them to emerge in changeable fields like social media through the discovery process (Naeem et al., 2023). This can be seen in investigations of influencer content, where TA reveals the underlying ideologies, such as empowerment in health advocacy or cautionary stories in financial warnings (Fife & Gossner, 2024). By identifying both latent and semantic themes, one can not only discover

what is said, but also the socio-cultural implications. At scale, however, challenges arise; manual TA is costly (Christou, 2022), so software-assisted coding has been developed to ensure rigor without sacrificing efficiency, especially when dealing with thousands of posts. The use of TA on social media emphasizes its application in narrowing the qualitative richness to meet the requirements of digital data analysis.

Finfluencers and Financial Narratives on Social Media

Finfluencers, or financial influencers, have become a significant figure in the digital sphere, combining education and entertainment (also known as edutainers) and spreading their financial tips through social media networks such as Instagram (Hayes & Ben-Shmuel, 2024). These influencers utilize engaging formats, such as short videos, infographics, and personal stories, to make finance more accessible, particularly to millennials and Gen Z, who seek a more relatable and approachable alternative to traditional advisory services (Rajkumar, 2025). The study conducted by Hii and Ong (2025) suggests that finfluencers can significantly contribute to financial socialization, as they simplify complex financial issues, making them more accessible and comprehensible, thereby influencing financial decision-making, such as budgeting and investing (Hii & Ong, 2025). In addition it was discussed that, on Instagram, finfluencers can be framed as financial literacy sellers since they engage with their subscribers, who are motivated to create financial-building projects despite financial uncertainty.

Empowerment, frugality, and risk mitigation are common themes in trending content of finfluencer (World Economic Forum, 2025). Empowerment narratives often portray finance as a path to freedom, and the idea of paying off debt or generating passive income is popular among younger

generations who struggle with student debt and the gig economy. Minimalism and wise spending are promoted as the primary components of frugality, which is complemented by the introduction of sustainable lifestyle principles that counteract the forces of consumerism (Ghadafi & Andriotis, 2025). The risk discourses, on their part, rely on the context, with some influencers advocating for risk aversion, including diversified portfolios, while others glorify risky sectors of operation, such as cryptocurrency trading (Govindarajan et al., 2025). The issue of gender disparities also emerges in the literature, where researchers have found that women tend to consult more when it comes to issues of security and long-term planning. In contrast, men often use aggressive investment methods due to social norms surrounding financial advice-seeking behavior (Mikelionytė & Lezgovko, 2021). These stories not only affect the decisions of individuals but also shape the financial culture of people involved.

Despite the proliferation of the content of finfluencers, critical loopholes are found in scale-based, recent works, primarily due to the dynamics of data volume and platforms (Mölders et al., 2025). Early studies often used a small number of participants or focused on TikTok because Instagram is a visual-aided platform where photos are seamlessly integrated with narratives. Recent regulatory issues underscore the importance of comprehensive surveys to track the trends and impact of these concerns on vulnerable populations (Okorie et al., 2024). Such gaps become especially notable considering how social media algorithms give more visibility to specific communicational choices. It is therefore essential to consider how these gaps can be minimized and what risks can be avoided, thereby leveraging the opportunities presented by the financial empowerment of influencers.

Use of LLMs in Thematic Analysis and Justification

In qualitative research methods, thematic analysis (TA) has been utilized in conjunction with large language models (LLMs) to provide a primary coding and theme extraction option for massive text volumes (Wong et al., 2025). Experiments with the models had been conducted in previous times to aid in identifying patterns in the interview transcripts. Interview transcripts are used to train the models that generate initial codes, which the researchers refine in subsequent runs (Hayes, 2025). LLMs approach is beneficial in situations where data is voluminous, noisy, and requires processing (Chiarello et al., 2024). They have also been applied to thematic clustering in educational research (e.g., to research student feedback or online discussions on a scale that could otherwise be impractical to explore manually) (Hayes, 2025; Wong et al., 2025). The applications emphasize the power of LLMs to recognize regularities, allowing them to address linguistic nuances and contextual equivalents, thereby substantiating qualitative results.

In this analysis, the application of LLM integration was guided by a clear and logical procedure according to the scheme proposed by Naeem et al. (2025), which introduces a systematic approach to using ChatGPT as a component of thematic analysis. This model enables reflexivity, replicability, and auditability by providing detailed descriptions of prompts, iterations, and code choices. The analysis made the LLM's output traceable and reproducible. The interpretation process is under the researcher's control through both fully controlled measures. Compliance with the recommendations suggested by Naeem et al. (2025) contributed to an increased level of methodological transparency and adherence to the current qualitative AI-supportive standards.

The use of LLMs in exhaustive exploratory TA can be a valid idea for several reasons. With fluent theme development, LLM promotes human reflexivity by prompting and refining (Vikan et al., 2025). They are efficient at handling large volumes of data, including the 22.854 posts used in this analysis, which would be impractical to code by individual human coders due to time and cognitive load limits. LLMs offer consistency in annotations per batch, thereby minimizing variability and bias due to human fatigue, and tracked outputs allow reproducibility. Human control remains the key, and researchers study the LLM-generated reports to confirm and refine themes, thereby ensuring interpretative richness (Vikan et al., 2025). Its benefits include affordability over employing numerous coders, scalability for large datasets, and the elimination of errors due to automated cross-checking.

Ethically, the decision-making of humans is not substituted by LLMs but is improved, which aligns with the hybrid TA approaches, according to which AI and researchers should work collaboratively (Bengani, 2025). Bias in algorithms can be mitigated through various training resources and a distinct call for supporting the prompting phenomenon, whereas privacy safeguards in data processing can ensure an ethical approach (Wei et al., 2025). This incorporation then increases the applicability of TA to the current research problems of the digital age.

Research Design and Framework

An exploratory and descriptive research design is employed in the present study, which is grounded in the approach of reflexive thematic analysis (TA) to identify patterns and themes in qualitative data from Instagram finfluencer captions. Reflexive TA, as described by Braun and Clarke, offers a flexible yet stringent approach to qualitative analysis, positioning the researcher as an active

participant in the reflexive interpretation of data (Politz, 2024). The structure is based on the six steps outlined in their seminal works: (1) becoming familiar with the data, including immersion in the captions to get a sense of what is in the data in general terms; (2) generating initial codes or keywords that articulate salient features; (3) searching the themes by clustering the codes into potential patterns; (4) reviewing and refining the themes to achieve coherence and relevance; (5) defining and naming the themes having clear descriptions; (6) production of report, which is putting the themes in a narrative description. This is a cyclic procedure, by nature, subject to constant reflection and modification as new understanding is gained through the data.

Since the study is exploratory, reflexive TA is quite suitable, since it does not focus on instilling categories in the data; instead, it involves the creation of themes based on the data (Terry & Hayfield, 2021). This is consistent with a constructionist epistemology, in which themes are viewed as actively constructed during the analytic process, and in which themes are more widely represented in terms of socio-cultural narratives in financial discourse (Burns et al., 2022). Nevertheless, due to the large size of the dataset, comprising 22.854 posts, a large language model was incorporated into the traditional method of human coding. In particular, the TA phases were run exhaustively with the help of the Llama 3.1 8B model, which was deployed locally using Ollama framework. This adaptation preserves the reflexive nature of TA because it models human annotation by prompting the data (Gillings et al., 2024), which is structured by interacting with the LLM in small chunks and updating the results with those findings.

The novelty of this new design lies in the LLM's ability to simulate human reflexivity by tracking external documents in real-time. Unlike a static automated tool, the LLM was prompted to refer to

and update an emerging external document containing accruing keywords, codes, and themes derived from previous batches, thereby generating a dynamic and iterative analysis (De Paoli & Mathis, 2024). This kind of methodology not only is scalable to a larger degree, but also assists in enhancing the richness of the exploratory findings, allowing the creation of a detailed map of the emergent discourses in finfluencer material. The synthesis of conventional TA ideas and AI advancements in the design can offer descriptive richness, as well as new methodological practices involving large qualitative datasets.

Data Collection

The dataset used in this study was obtained through the Instagram Finfluencers Dataset, a custom-designed compilation of 22.854 captions of public Instagram posts by prominent financial influencers. Captions were chosen as the main unit of analysis because of their rich textual nature, which tends to condense major messages, pieces of advice, and stories that support visual information. The detailed process through with the dataset was built is described in the section *"BUILDING THE DATASET"*, that will follow this one.

The primary concern during the process was ethical. All the information was obtained through publicly available posts, without violating the privacy of users by not using private accounts. Identity was hidden by deleting identifiable items, such as usernames and timestamps, and the dataset was converted into a de-identified text snippet. This practice reduces the risk of harm to the influencers, including reputational risk of misinterpretation, and complies with ethical principles in conducting social media research, including do-no-harm principles and data minimization.

Personal information was not collected, and the database was stored securely locally only for the time required by the analysis functional to the study.

Data Analysis Procedure

To process such a vast amount of data, the 22.854 captions were broken into batches of 200 posts, which made systematizing easier and replicated the gradual reading of a human TA. This batching technique enabled the application of the LLM to the dataset in a controlled manner, which guaranteed computational efficiency and refinement.

The TA, facilitated with the help of the LLM, was implemented in accordance with the stages of the scheme, as outlined by Braun and Clarke (2022), and taught using specially designed prompts to the Llama 3.1 8B model. In each batch, this would be done through familiarization, during which the LLM would be asked to read and summarize the captions to identify general patterns. Initial keywords with which the semantic elements were determined included, but were not limited to, “Financial Feminists” or “Debt-Free Journey”. On these, coding or conceptions were formed (e.g., Feminist Pedagogy or Behavioral Economics Tips), in which similar ideas were grouped into interpretative units. Themes in the LLM were subsequently constructed or refined, e.g., creating T1: Women-First Money Empowerment. This is achieved by collating codes and verifying their consistency.

Incremental annotation, in which the output of each batch was recorded in an external document that served as a reflexive log, was also a key innovation. This document was initialized with baseline instructions, added keywords, codes, and themes, and presented to the LLM in repeated

batches for reference. The LLM was asked to combine redundant keywords, such as “Girl Boss Finance,” with existing ones or refine theme definitions with new information. This would mimic human reflexivity, a process of reading, annotating, and modifying findings that allow for emergent modifications. The researcher reviewed more specific reports generated by the LLM (summaries, code mappings, and theme justifications) after every batch and made adjustments to address inconsistencies or initiate further iterations.

The utility of the method is demonstrated in terms of its benefits: scalability has been achieved, as the LLM enables exhaustive analysis of 22.854 posts, which would have overwhelmed the manual process; objectivity, since the LLM lowers the subjective biases that would otherwise be present in the process; replicability, since the audit trail in the external document can be used to verify; and resourcefulness, where AI was leveraged to conduct comprehensive TA with small teams. Additionally, it reflects responsiveness to data patterns: recognizable (data-based themes), reciprocal (iterative dialogue between batches), responsive (adaptations to new insights), and resourceful (efficient use of technology), which makes it a strong exploratory analysis.

Rigor and Trustworthiness

To maintain rigor in this LLM-assisted TA, several measures were implemented to ensure trustworthiness is maintained. Dependability was also ensured in a comprehensive audit trail, the external document, and batch report, which record all the steps in theme development in a transparent and replicable way. To enhance credibility, the prompts of the LLM were aligned with the TA standards (the importance of reflexive interpretation), and human control was employed to authenticate the outputs against the raw data samples. The themes facilitate transferability due to

their potential generalizability to the broader context of social media finance, and the detailed descriptions enable readers to evaluate their applicability.

Limitations include the potential for hallucinations about the LLM, which can be mitigated by cross-verifying outputs against subsets of data and refining prompts to ensure validity and accuracy. Overall, these methods contribute to the integrity of the study, in search of a balance between innovative and reliable methodological research.

Tools and Ethical Considerations

The primary analysis tool was Ollama, an open-source local LLM deployment platform, which utilizes a 8B Llama model and is deployed on a local GPU platform to ensure privacy and control over computation (IBM Technology, 2025). Ethical implications were also considered in the analysis of data, ensuring that all data is handled locally and no cloud services are used, which would avoid data exposure to third parties. The adherence to the ethical guidelines was crucial in making sure that AI was used responsibly in social research.

Results: Thematic Map of Finfluencer Narratives

The synthesized findings inform this section of the coding and theme development process applied to Instagram finfluencer content, based on the reflexive approach of thematic analysis as developed by Braun and Clarke (2022). Thirty-six themes were identified, each representing a unique yet related story about financial discourse on the platform. The themes are organized into seven broader functional clusters, and they denote the pedagogical and sociocultural dimensions of influencer communication. All the themes, their keywords, and coding summaries are listed in

Appendix A. The synthesis below explains each cluster separately and how they form the thematic map of finfluencer narratives.

Foundations of Personal Finance

The Foundations of Personal Finance cluster frames the finfluencer space with a series of sequential literacies, including budgeting (T2), emergency funds (T3), debt management (T4), credit health (T5), and retirement and investing practices (T7 and T8). Budgeting is presented as the optimization of workflow, rather than deprivation, with a focus on automation and so-called cash flow systems that liberate the mind (T2). The content of emergency funds (T3) redefines liquidity in terms of self-protection and emotional stability, and debt-reduction themes (T4) combine behavior psychology and mathematical reasoning to ensure compliance. Credit health (T5) categorizes borrowing as infrastructure, focusing on literacy regarding the reports and their use, rather than score hacks.

These pillars result in the retirement and investing pedagogy (T7), in which designers reduce complex systems to order-of-operations routines: capturing the employer match, and employing a passive and long-term investing strategy with index funds and dollar-cost averaging. This group is effective in operationalizing financial literacy as a set of small, repeat actions. It is a conversational, low-key, and algorithmically shareable discourse that suggests how finfluencers have redirected their pedagogical interests towards applied, identity-based learning, rather than abstract knowledge.

Income and Growth Pathways

Career Capital (T12), Side Hustles and Entrepreneurship (T13), and Small-Business Finance (T22) suggest that generating income is a technical and narrative process. Finfluencers train career capital on language and evidence resume optimization, scripts to negotiate, and literacy, bridging the divide between self-worth and market quantification. Autonomy and accessibility are viewed as important in a side-hustle discourse: at first, they are minor, but are added later and appreciated based on the work, not the hours. Business-finance content (T22) expands on this logic by discussing cost structures, taxes, and compliance, and advising creators and entrepreneurs to work on a cash flow basis. All these themes make personal finance not just about saving, but about earning power, wherein it is argued that financial security needs to be not only frugal but diversified and organized through income.

Identity and Empowerment

The Identity and Empowerment themes assume mobilization of belonging as it is promoted by finfluencers against traditional exclusion in finance. Women-First Money Empowerment (T1) reorients money education as a collective identity process, and focuses on using collective-centered language (builder girlies, financial feminists) and anti-shame pedagogy as an entry approach to decrease psychological barriers to entry. Women of Color and First-Gen Wealth Narratives (T23) add to these by making money education culturally resonant and representationally transparent, thus reducing psychological barriers to entry. Then, Advocacy and Civic Issues (T16) ties personal finance to structural inequities, connecting issues like healthcare, childcare, and wage gaps to financial agencies. Throughout these stories, money serves as both a source of independence and a means of social expression. This empowerment, rooted in identity, aligns with feminist financial

arguments that emphasize the importance of safety, boundaries, and solidarity as essential prerequisites for economic engagement.

Social and Relational Dimensions

Finfluencers are making a move to incorporate financial literacy into their personal and family relationships. Money and Relationships (T17) themes normalize the openness of finances, prenuptial agreements, and role negotiation in couples, allowing them to discuss issues rather than contend with them. Parenting and Intergenerational Planning (T18) is a longer-term project that showcases wealth as a family initiative, with children saving money and receiving early credit education. Health × Money (T24) is a wellness economics study that relates preventive care, fitness, and medical expenditure to long-term stability. When combined, these themes of relationships broaden the lexicon of financial advice to include the collective by introducing money as a collective governance model that facilitates emotional well-being and lineage.

Meta-Themes and Adherence

The Meta-Themes and Adherence cluster (T19, T15, T14) describes the behavioral and infrastructural supports that maintain engagement. The contents of Values, Mindset, and Mental Health (T19) emphasize the importance of discipline over motivation and advocate that systems and accountability partners are the remedy for burnout. These mindsets are operationalized by CTA and Community Offers (T15) structured calls-to-action, live workshops, and membership cohorts that convert attention to practice. The Creator Economy and Audience Growth (T14) simultaneously define the meta-layer of finfluencer sustainability: the acquisition of owned audiences through email lists and live education, thereby reducing reliance on algorithms. The idea

of behavioral design unites these themes, the concept that the desired transformation in finances is not inspired but instead perpetuated by consistency (Politz, 2024).

Advanced and Market-Oriented Themes

The Advanced/Market-Oriented group refers to an even newer frontier of finfluencer pedagogy. Trader Education (T31) documents educational programs that emphasize execution skills, such as risk management, stop-loss habits, and position sizing, as part of entry-level mentorship programs. Fintech Infrastructure (T32) rethinks the structure of the new market, providing a rationale for the three instruments of access, namely, custody, stablecoins, and prediction contracts, as tools of access rather than speculation. Collectively, these themes demonstrate how finfluencers transform professional finance into modular, consumer-friendly learning platforms, without being too predatory.

Promotion, Marketization, and Financial Subcultures

This cluster focuses on the intersection of financial education, online subcultures, and digital marketing. Promotion of Products, Events, and Community (T34) also redefines learning as an ongoing process, utilizing postings, polls, and live events that serve as funnels in the process of winning and retaining audiences. Newsletters, Telegram groups, and link-in-bio setups are the foundation of a perpetually looping community that Finfluencers can build. As explained in Crypto: Between Institutionalization and Hype (T35), the digital finance industry is a hybrid organization that is both credible and speculative. While memecoin volatility might be based upon hype, educating people about risk and discipline, the institutional adoption model, through ETFs and the rails of custody, is the virtual opposite. Finally, the third research area is Mindset, Identity, and the

Trader Persona (T36), which pertains to the cultural psychology of trade. Stoic mottoes and humor make discipline a matter of social belonging, and memes and rituals reinforce resilience. Taken together, these subcultural forms bring together education, entertainment, and marketing to form a self-sustaining ecosystem of engagement.

Themes in Light of the Literature

The findings of this thematic analysis describe how Instagram influencers are transforming financial education into a comprehensive ecosystem that encompasses literacy, community, and behavioral design. To contextualize these findings, this discussion matches the thirty-six emergent themes with related academic literature in feminist economics, behavioral finance, psychology, sociology, and media studies. This synthesis highlights that finfluencers do not passively reproduce preexisting ideas about financial behavior; they creatively produce them within the culture of digitality. Personal finance has been transformed into a participatory, emotive, and social networked practice.

Feminist economics align with the women-first content produced by finfluencers (T1), as inclusion and relationship-based learning have the potential to decrease gender gaps in financial literacy (Sundarasen et al., 2023). They make finance a group activity, not an individual one, through the deployment of anti-shame and empowerment language. Behavioral finance research on automation and choice architecture, aided by budget systems and cash flow systems (T2), seeks to simplify the decision-making process and foster habits (Goud et al., 2024). Emergency fund advice (T3) is consistent with precautionary savings theory, which suggests that liquidity buffers are beneficial as they foster resilience and confidence in financial circumstances (Babiarz & Robb, 2023). The debt

management and credit repair (T4) theme suggests that there is evidence that small steps and tips for constructing practical repayment habits help enhance compliance and reduce anxiety (Babiarz & Robb, 2023). Together, these themes reveal how finfluencers frame financial education within a framework of empowerment, emotional nurturing, and pragmatic safety, incorporating elements such as behavioral design, feminist empowerment, and practical safety.

Some of the themes that illustrate the effects of the gig economy on financial identity include income diversification (T5), side hustles (T6), and entrepreneurship (T7). The current body of research on self-employment associates diversification with autonomy and economic security, but this comes at the cost of burnout and precarity (Holloway & Pimlott-Wilson, 2021). To mitigate this conflict, persuaders introduce additional revenue streams in the form of empowerment tactics, alongside boundary-setting memories from feminist labor studies that value agency in flexibility (Sundarasen et al., 2023). The fact that classical finance knowledge is extended through investing in education (T8) implies that index-based, long-term investment is more desirable then active management (Subagio et al., 2021). Finfluencers then translate such evidence into similar storylines, such as community challenges and visual progress indicators, to encourage the community to be patient. Integrating entrepreneurial spirit and investment discipline, these themes create financial independence not as speculation but as a strategic, diversified engagement in accordance with the principles of sustainable wealth-building and behavioral compliance.

Themes of financial planning and goal-setting (T9) are similar to the goal-setting theory of Ang (2024), which emphasizes the importance of clarity and feedback in achieving enduring progress. This is applied by finfluencers who employ visual trackers and milestones-dependent budgeting,

which converts abstract goals into habitual ones that prove to be accomplishable. Protection and fraud awareness (T10) aligns with consumer-behavior research, which indicates that procedural knowledge and routine vigilance contribute to a decreased risk (Crasta, 2024). The economic security systems are enhanced by literacy in insurance (T11), which helps stabilize household finances (Crasta, 2024). Goud et al. (2024) provide evidence that long-term financial health can be predicted by credit building and scoring (T12). Both of these themes demonstrate the extent to which the idea of finfluencers incorporates the concept of a safety net, systematic planning, and a knowledge support system in everyday communication, as well as the way in which complex institutional systems are made palatable through their repetitive messages and emotional support.

The themes of accountability group (T13) and habit tracking (T14) are echoed by social learning theory, which posits that peer modeling and feedback loops maintain commitment (McLeod, 2025). The rituals and the challenge group of finfluencers represent this type of participatory pedagogy. Community-based accountability is explored in the creator economy (T15), where individual brands act as pedagogical centers. A study on influencer education reveals that authenticity and transparency increase trust and compliance (Baghel, 2024). Advocacy and civic engagement (T16) combine microfinance with structural literacy. Björklund (2021) refers to this as economic citizenship, in which financial ability is connected to systemic justice. With the matching of budgetary and civil matters, such as equitable pay, finfluencers convert personal education into general awareness. Throughout those four themes, education is bound up in entanglements: learning is not preserved by authority but through collective culture, where communities co-produce accountability, empowerment, and ethical involvement in the wider economy.

The social learning theory argues for group accountability (T13) and habit tracking (T14) themes, providing commitment through modeling peers' behavior and feedback loops (McLeod, 2025). Rituals and challenge groups of finfluencers are examples of participatory pedagogy. One significant component of the creator economy is community-based responsibilities (T15), where individual brands act as learning institutions. Authenticity and transparency are found to increase trust and compliance in the study of influencer education (Khalfallah & Keller, 2025). Combined with structural literacy, advocacy, and civic engagement, microfinance is a powerful tool (T16). Björklund (2021) refers to such economic citizenship, in which systemic justice is connected with financial capacity. When the budgetary and civil considerations are equalized, as in equitable remuneration, the finfluencers will convert their personal knowledge into a general know-how (Fornero & Lo Prete, 2023). The study of these four themes is confounded: learning is not something that exists and is preserved by authority, but rather it is co-produced by the community, where communities co-produce responsibility, empowerment, and moral involvement in the larger economy.

Safety and fraud prevention (T21) is a topic that follows the discussion about protection and the research conducted by Sabrin et al. (2025), who also note that repetitive reminders and transparency can support ethical behavior. Through their narratives, finfluencers make vigilance and agency normal. Behavioral studies on emergency preparedness and resilience (T22) reveals that precommitment leads to reduced exposure to crisis (Lillywhite & Wolbring, 2022). Sustainable consumerism content on sustainable consumption (T23) overlaps with environmental economics, which connects personal finance with ethical consumerism (Tomșa et al., 2021). Finally, intergenerational wealth (T24) is also reflected in sociological works that suggest that exposure

and modelling of families at an early age affect economic mobility (Subagio et al., 2021). The combination of all these themes demonstrates that finfluencers position financial literacy within the framework of moral and intertemporal reality, emphasizing the significance of sustainability, foresight, and ethics. Their online pedagogy is a composite of personal preparedness and system realization, self-defense, and social duty, as the economic benefits are complementary.

The approach of Finfluencers toward digital tools and applications (T25) is closely linked to the knowledge of online and mobile finance literature, which examines usability and perceived control as triggers to interaction (Kumar, 2025). Tax compliance education (T26) underlies the policy hypothesis, which states that simplified explanations can enhance the accuracy of filing (Surugiu et al., 2025). Themes of business financing and lending (T27) support one of the key findings of entrepreneurial finance studies, which indicate that mentorship and literacy are predictors of success (Kicova et al., 2025). Bargaining and pay equity (T28) reflect feminist labor studies in which transparency is viewed as one of the few mechanisms that can close wage differences (ILO, 2024). Collectively, these themes articulate how finfluencers expand traditional financial education into broader structural domains, legal, entrepreneurial, and professional, for individuals, aiming to strengthen them to better engage with systems that have historically excluded them. Their pedagogy intertwines literacy with agency, handing followers the tools to view negotiation, compliance, and digital mastery as civic skills.

Themes related to digital marketing awareness (T29) and online brand-building (T30) align with media studies that show creators in participatory cultures combining education and entrepreneurship (Mammassis, 2025). Influencers view personal brands as a mechanism of trust,

where attention is translated into a continuation of learning. Trading education (T31) is based on behavioral finance evidence advising against excessive trading and overconfidence (Trinugroho & Sembel, 2011), which finfluencers combat with rules-based education, including journaling and discipline. Fintech literacy (T32) helps Yu et al. (2024) to argue that unlike in the situation with digital platforms, where barriers to access are lowered, risk assessment requires an informed evaluation. Finfluencers across these themes are striking a balance between accessibility and caution, through personal engagement and technical teachings. The finfluencers, within the hybrid education-as-community model, transform transactional learning into continuous mentorship.

Corporate ethics research is directly linked with ethical marketing and transparency (T33), where higher transparency compared to the usual conditions is found to enhance consumer confidence (Ali et al., 2025). Products, events, and community advertisements (T34) are in line with the community flywheel in which learning and retention are conditioned by the engagement loops estimated by Hoffman and Fodor (2010). This tension between institutionalization and speculation is articulated in crypto education (T35). Kumar (2025) documents the technological maturity progress of this technology, and Sabrin et al. (2025) demonstrate narrative contagion. Finfluencers are also operating in the context of this two-sided risk-literacy model (Issac & Seranmadevi, 2024), which will make digital citizens more aware. Trader behavior and attitude (T36) is also related to Bandura's (1986) social-learning theory and to Baghel's (2024) investor psychology familiarity, since it is found that identity and humor offer a persistence system. Together, these themes underscore how finfluencers can blend ethics, innovation, and cultural identity to educate people on financial matters while promoting a culture of trust, storytelling, and solidarity.

Conclusion

Using an LLM to analyze a dataset of 22.854 posts by finfluencers on Instagram, this study identified 36 themes in seven clusters: Foundations of Personal Finance, Income Growth, Identity Empowerment, Social Dimensions, Meta-Themes, and Advanced Topics. The findings indicate that finfluencers play a part in the formation of financial education as a participatory-behavioral ecosystem that involves a dialectic of literacy and identity making as well as community-oriented responsibility. This methodological strategy was effective enough to be used to argue that the scale of LLMs can be expanded to a full-fledged qualitative study of large amounts of data. The themes are aligned with recent behavioral economics and financial literacy theory, and they provide new insights into the role of social identity and feminist pedagogy in contemporary financial discourse.

Some of the limitations of this study are using a single LLM, which, though it has potential, may be biased in terms of model performance. Additionally, the emphasis on Instagram captioning is inapplicable to other audiovisual platforms, such as TikTok. Future scholars are advised to seek hybrid methods of balancing human-LLM methods and thematic analysis in a bid to reach balance between scalability and interpretive richness. They ought to also conduct longitudinal research on how the story of the finfluencer developed. In practice, the insights are relevant in respect of the further promotion of more effective financial literacy programs and regulating regimes, to understand the significant role of finfluencers and consider the opportunities and risks of the new media.

BUILDING THE DATASET

This section is dedicated to briefly outline the methodology used to build the dataset that was used in the previous thematic analysis and that will be used, in an enriched version, in the analysis in the last section.

Scraping Ethical and Legal Compliance

A custom web crawler was developed to navigate Instagram simulating an ordinary user activity. Compliance with legal and ethical principles discussed in the previous dedicated section was achieved by observing the following precautions in accordance with GDPR principles and the discussed ethical basis for web scraping implementation:

- No copyright content was collected to avoid copyright infringement.
- No paywall or subscription-based service was bypassed.
- Even if Robot.txt documents are not legally binding a workaround to protect integrity of the website was implemented. In particular, the crawler did not take advantage of scalability or parallelization of processes and instead closely simulated real user activity by pausing on each link for 6.0 second before resuming activity, so to avoid overloading the website more than a common user, also in accordance with the principles of "polite scraping".
- Only profiles that were public at the time were scraped.
- No information that can result in reidentification risk was collected. Usernames, timestamps, personal links were ignored and never collected.
- Textual data was collected and stored only for the purposes of research and for the time necessary to execute the analysis.

- Where the analysis required post images, no image was stored on a local machine. The relevant information was extracted “on the go” through static links, and no information extracted from the images pertained to profile identity or copyrighted material.
- Legitimate interest is constituted by academic research and public interest in the results.

Defining the Scope of the Scraper

The universe of the posts to scrape was first defined broadly by using hashtags and keywords that are common in the financial social media context and were used with a loose correspondence policy to maximize the number of eligible cases. Some of them are, but not limited to:

- Keywords: "financial influencers", "personal finance experts", "money advice", "financial literacy", "wealth-building tips", "financial independence", "financial education", "financial coach", "financial mentor", "trading signals", "stock tips", "forex trading", "crypto trading", "day trading", "swing trading", "trading strategies", "trading courses", "trading mentors", "stock market alerts"….
- Hashtags: #FinancialLiteracy, #MoneyTips, #WealthBuilding, #FinancialFreedom, #PersonalFinance, #FinancialEducation, #MoneyManagement, #InvestingTips, #FinancialCoach, #SmartMoney, #TradingSignals, #StockMarket, #ForexTrading, #CryptoTrading, #DayTrading, #SwingTrading, #TradingStrategies, #TradingTips, #StockTips, #TradingCommunity…

Only posts from 01-01-2024 to 30-09-2025 were selected and ordered starting from the ones with a higher absolute number of interactions. This first search was aimed only at collecting static links

for the posts and resulted in a first version of the dataset stopping the collection at 96.539 records due to both dataset manageability and clear degradation in relevance and number of interactions.

Then a multiple filter approach was adopted. Only links of posts that had more than 100 likes, at least 1 comment and a non-empty caption were kept, as well as only posts belonging to profiles with more than 10.000 followers, which was the threshold that until recently allowed Instagram users to paste working links in stories. Then an additional semi-manual filtering step followed. To guarantee relevance of the posts before proceeding with the scraping of the needed information, a list of the profiles was redacted and only profiles that fulfilled certain conditions were kept. To be eligible, profiles had to show finance-related keywords in their profile description, had to have published content at least once in 2025 and had to show out-of-platform web presence. These precautions were implemented to guarantee recency and relevance, since profiles that are confined to Instagram cannot attest any impact outside the platform.

Scraping of the posts from the selected profiles, in the same interval of time and applying the same post filters described above, resulted in the final 22.854 records dataset.

Object Detection Models Training and Data Enrichment

The visual information enrichment process was conducted using YOLOv11. The model was trained to detect the following classes: person, suit, tie, car, boat, banknote, coin, trading chart, while the added class money is given as presence of either coin or banknote. To train the model a custom

dataset was assembled by collecting annotated images with the relevant class occurring one or more times. 5000 images for person, 5000 for car, 3000 for tie and 3000 for boat were collected from COCO dataset (Lin et al., 2014), while for other classes 900 images were collected for each one from publicly available sources. The resulting dataset was splitted in train, validate and test subsets respectively accounting for 70, 20 and 10 percent of the original dataset. Default training parameters from Ultralytics configuration (Ultralytics, 2025) delivered satisfactory performance results. The model was applied to the images while the scraper passed from one link to the other allowing for processing images without storing. Contextually the dataset was updated in real time by registering the detected classes each as a different binary variable.

The Issue of Reproducibility

While each component of the process described above is easily reproducible, exact reproducibility of the dataset is not guaranteed. The reasons for this issue are multiple, but two are the most relevant. The first reason is technical. Scraping process is not guaranteed to always correctly fetch each web page correctly. Even if approached with polite scraping, many additional non-intuitive issues are to be considered to navigate this social media smoothly via a crawler. That means that two different runs can yield slightly different results. The second reason is way more relevant and has to do with the nature of the topic under study. Social media posts are not a static reality. Content creators manage and modify regularly their past content. The older the dataset, the higher the risk the processed posts got removed, modified or the profile switched to private status. This risk is greater in a fast-paced environment like the one finfluencers move in. By the time collection of the data for the fully filtered dataset was completed (in about 4 days) some posts processed on the first day were already unavailable. Then the way the dataset in use here is to be understood is as a

representative snapshot. On the other hand, precautions about recency and relevance of the finfluencers' profiles involved should greatly reduce this issue and considerably enhance the intertemporal reliability of the analysis that will follow.

BUILDING A MODEL TO PREDICT TEXTUAL CONTENT USING VISUAL CUES

In this section a predictive model will be developed. As it was already explained in the introduction it will exploit the visual/symbolic elements present in the images of the posts to infer with reasonable probability the textual contents of the associated post. The dataset in use in this analysis is derived from the *Instagram Finfluencer Dataset*. In particular the original 22.854 posts dataset was enriched both by executing a LLM-assisted remapping of the themes discovered in the previous section to each post and by implementing an object detection model, YOLOv11, specifically trained to detect object classes that hold symbolical significance in the domain of finfluencers in the posts images (person, suit, tie, car, boat, banknote, coin, trading chart).

Choosing Visual Cues

In this context such objects are to be understood as the props that influencer regularly use to construct their image and identity (Soto-Vásquez & Jimenez, 2022). Finfluencers are a special case of influencers; as such they also must find balance between different aspects of their personal brand to convey trustworthiness and authenticity (Kim & Kim, 2023); which are also the necessary conditions for monetization. The rationale for searching for symbolically relevant props in images from finfluencer promotional activity simply takes note of the facts that social media communication is a multimodal enterprise and the construction of credibility by influencers, and as such by finfluencers, is also built upon implicit communicational expediencies (Meer & Staubach, 2020). The choice of specific props, or object classes in the scope of the object detection task, was derived deductively by the examination of 922 videos scraped from TikTok and selected by relevance and volume of interactions to ensure both coherence with the thematic domain and

proven communicational effectiveness. The reasons for using a different social platform than Instagram to operate this selection are multiple. First, TikTok is a video-based social platform and as such offers a considerably denser stream of potential props and visual cues to search for and examine. Second, if the chosen props really hold symbolic relevance, then their effectiveness as predictive elements for the thematic context should carry from one social platform to the other. In fact, eventual confirmation of cross-platform relevance constitutes an additional confirmation that the chosen props hold symbolic significance in the finfluencer domain as a whole and are not specific to sub-domains.

Dataset Enrichment

The remapping of the themes to the posts was obtained by instructing a Llama 3.1 8B model to search out elements of each theme in the caption of a post. The model had access to the table in Appendix B and to more detailed descriptions of the themes derived from previous section as external documents to exploit via Retrieval-Augmented Generation. The task was iterated on each caption for each theme to guarantee exhaustiveness and binary variable coded the attribution of themes to captions. Detection of the relevant objects/classes was also coded via binary variables registering presence or absence of each object class.

Subsequently only the records of the dataset that presented both eligible values for the features (the detected object classes) and the targets (remapped themes) were kept. This guaranteed a fully populated dataset that is susceptible to effective analysis but reduced the usable records from

22.854 to 10.492. The following step was implementing a predictive model that used the visual elements in the images to infer the topic in the captions.

Model Choice and Implementation

To carry out this task a random forest model will be used. Random forest models are relatively new, but far from exotic for social sciences researchers. They already proved their effectiveness and utility in a range of research fields heterogeneous enough to highlight their versatility, such as behavioral psychology, demography and political science (Best et al., 2021; Fife & D'Onofrio, 2023; Jones & Linder, 2015). The choice of this model is motivated by some advantages that it has compared to traditional linear models (Schonlau & Zou, 2020; Marchese Robinson et al., 2017). Random forests naturally capture nonlinear relations and interactions because each base learner is a different decision tree. Aggregating many trees reduces variance and yields predictions that are more stable than a single tree. The algorithm combats overfitting through two layers of randomness: bootstrap sampling of cases and random subsetting of variables at each split, which decorrelate trees. This technique is effective also when the number of predictors rivals or exceeds the number of observations, requires minimal preprocessing (no scaling and few distributional assumptions), tolerates multicollinearity by letting trees select among correlated cues, and accommodates mixed data types in a single model. Out-of-bag evaluation provides an internal cross-validation signal without a separate holdout. Finally, it offers variable-importance diagnostics, both permutation-based and impurity-based, that help interpret which inputs are the most relevant.

Procedure Implemented

K-fold cross-validation is employed with shuffling and a fixed random seed. The number of folds is determined dynamically as the minimum of five and the minority-class count, with the explicit requirement of at least five positives per fold; this avoids degenerate splits when positives are scarce while keeping class balance within folds.

Within each cross-validation split, a model is trained on the training portion using 500 trees, unrestricted depth, and a minimum leaf size of five observations to regularize the trees. Class imbalance is addressed through automatically weighing each class inversely to its prevalence, so the minority class contributes more to the splitting/loss criterion. In this way class imbalance is mitigated during training. The model outputs class probabilities on the validation portion only; those out-of-fold scores are the sole basis for performance estimation and feature-relevance diagnostics, so no training information leaks into evaluation.

A loose, nested grid search was executed inside each outer cross-validation fold. The inner search tuned the forest over two sizes (250 and 500 trees), three depth settings (None, 10, 20), three values for the minimum number of samples required to split an internal node (2, 5, 10), and three values for the minimum number of samples required at a leaf (2, 5, 10).

How Feature Relevance is Computed

The primary relevance number is permutation importance measured on Average Precision. On the validation data of each fold, the baseline Average Precision is recorded; then, for each feature in turn, the column is randomly permuted in the validation set to neutralize its association with the target. This is repeated twenty times, and the resulting Average Precision values are averaged. The importance for that feature in that fold is the drop in Average Precision between the baseline and the permuted condition. The procedure is also repeated with Area Under the Receiver Operating Characteristic Curve as the scoring function to provide a diagnostic view that is less sensitive to class imbalance. Importances are then averaged across folds, and features are ranked by permutation-on-Average-Precision.

Performance Metrics and their Interpretation

Multiple fold-level metrics are computed and then averaged with associated standard deviations. The first is Area Under the Receiver Operating Characteristic Curve (ROC-AUC), which express the probability that a randomly chosen positive receives a higher score than a randomly chosen negative; a value of 0.5 indicates chance-level ranking, noting that AUC can look optimistic under severe imbalance. The second is Average Precision (AP), the area under the precision-recall curve, which focuses on the positive class and is better aligned with imbalanced scenarios. Because AP depends strongly on prevalence, also AP lift is reported, defined as mean AP divided by the base rate (positives divided by total n). An AP lift near 1.0 suggests no gain over a naive ranking. A permutation-importance value of 0.08 on Average Precision means that shuffling that feature reduces Average Precision by 0.08 points on average across folds and permutations; larger positive

drops indicate greater contribution, values near zero indicate negligible effect, and negative values typically reflect noise or instability rather than genuine predictive value. Gini importance is unitless, sums to one within a model, and can be biased toward variables with many split points, so it is treated as corroborative rather than primary.

Design Choices that Support Robustness

All evaluation and permutation steps operate strictly on out-of-fold data to prevent optimistic bias. Stratified folds and class weighting mitigate the distortions of class imbalance. The dynamic determination of folds ensures that cross-validation is not attempted when the data cannot support it. The choice of adopting a loose grid for hyperparameter optimization allows for minimal adaptation of the model to the data without risking overfitting via fine-tuning. Repeating permutation twenty times per feature and averaging across folds reduces Monte-Carlo variance. The fixed random seed ensures full reproducibility.

Evaluating Results

Prediction for 27 out of 36 target themes was substantially unsuccessful. This is not unsurprising. Segmenting the dataset into 36 smaller chunks of variable size, even if there can overlap between them, since more than a theme can be mapped to the same caption, may obstructs the ability of the model of building predictions on a detectable and stable signal from the features. These difficulties are reinforced by the fact that the feature space was already sparse, and very few features contributed conjunctly to predictions. On the other hand, this also means that where modeling attempts were successful the validity of the prediction is to be regarded as intuitively stronger. In

contrast with the case of p-value, AUC and AP lift cannot rely on an ingrained and well-established convention to baptize a threshold of significance. Which value is to be regarded as satisfying is heavily dependent on the field of application. In this epistemologically challenging environment, an AUC > 0.60 is to be regarded at least as worth of notice, while values that approximate to 0.65 are already reliable. AUC > 0.70 is considerable a reliable prediction and a symptom of a clean and readable signal, while values above 0.80 are outstanding. In similar fashion, an AP lift > 1.0 is clearly insufficient, since it means that the prediction cannot discriminate between a signal and the general tendency of an imbalanced class, but AP lift > 1.50 would already suggest a relevant result. AP lift > 2.0 constitutes strong confirmation. In the next paragraphs relevant targets will be reported and discussed, but an exhaustive table of the random forest analysis results can be found in Appendix B.

Discussion of Empirical Results

Target predictions that are worth discussing pertain to themes number 1, 12, 13, 16, 31, 32, 34, 35 and 36. Performance of every target prediction will be briefly discussed and an intuitive account and understanding of the result will be attempted; as well as eventual confirmations from the established literature will be outlined.

1 - Women-First Money Empowerment (Financial Feminism)

This theme revolves around the new trend of women participation in the personal finance space. What is observed is the reframing of personal finance as a collective competence rather than an individual test, while focusing on women-specific issues. The anti-shame stance (“welcome,

Financial Feminists") and boundary talk ("say no to performative spending") lower emotional friction on women, and see education also as identity work, by for example calling out biases such as "the beauty tax".

With AUC = 0.649 ± 0.031 and AP lift = 1.77 the model performance is just satisfying. But what strikes as noticeable is that it is the only the class person, the most common class in the feature space, to do the heavy lifting, with a permutation AP (0.012) many times higher than the one of the second feature (0.003) for importance, suit. This suggests that the presence of the class in posts that expresses this theme is prominent enough to allow an educated guess. Intuitively speaking it seems that an explicitly progressive stance and pushing empowerment narratives really call for a face. Finfluencers that insist on these topics seem to prefer showing their person to communicate honesty and authenticity to an audience in search of a stronger personal connection. This is coherent with the fact that financial feminism implies reformulating financial issues also as a function of personal identity.

12 - Income Growth: Career Capital

In this case the discourse moves to a more impersonal tone. Here the focus insists on optimizing hirability and to manage layoff risk though cultivating personal value through competence (prove value, speak value, and collect value"). But also cultivating negotiation skills.

Coherently with the focus on job market and professionalization, the two most important features for the prediction task are suit (Perm-AP = 0.017) and tie (0.012), with the class “person” being substantially irrelevant (0.002). The intuitive alignment between visual cues and thematic scope starts to function as a validating principle of the analysis. While the present association is less interesting from an informative point of view than the previous one, a solid AUC of 0.675 ± 0.009 and an AP lift of 1.96 stand as confirmation of the validity of the chosen methodology.

13 - Side Hustles & Entrepreneurship

Unsurprisingly thematic analysis also captured content about side hustle and individual entrepreneurship culture. The narrative of this topic stands on small business advising, business literacy, and easy-to-start side activity such as dropshipping. The finfluencers that push these narratives want to transmit professionality and competence. Their personal image choice involves suit (0.129) and tie (0.123) with an importance considerably greater than in the case of *Income Growth*.

An AUC of 0.708 ± 0.025 and an AP lift of 2.93 confirm the narratives of side hustling, small business celebration and personal entrepreneurship as reliably detectable and easy to associate to a particular choice of visual props.

16 - Advocacy & Civic Issues via Money

The use of money to explicitly promote civil causes such as framing abortion as financial issue, donation matches, voter info, black wealth panels and community calls, also found its place between the themes that emerged. But despite an abundant presence of this topicamong the dataset records (14.65%) the evidence for ties with specific visual props is weak (AUC = 0.610 ± 0.010, AP lift 1.31). But that is no evidence that the theme itself is irrelevant to the finfluencer space. In fact, the contrary can be argued. Connections between finance and civil issues is an extremely rich topic, which happens to be transversal to many social media contexts. It can be hypothesized that tying such a wide too few visual elements is the wrong approach to tackle this particular topic. The Most probable cause for ineffective predictions in this case, despite the abundance of datapoints, is that the feature space delineated in the experiment is not rich enough to capture the theme. That should serve as a suggestion to study the topic more carefully and enrich the features the model can leverage.

31 - Trader Education & Execution

With Trader education & Execution a completely new thematic landscape is emerging. Finfluencers associated with this theme focus on the technical and practical aspects of retail trading. They educated their followers on terms like pips, stop-loss, candlestick entries, and promote mentorship programs and beginner-friendly workshops that regard becoming retail traders. These finfluencers make extensive use of trading charts, that they rely on to explain market events and illustrate educational content often in the conceptual framework of “technical analysis”. This peculiar

reliance on candlestick trading charts (0.244) was easily captured by the model and translated in very stable predictions (AUC = 0.821 ± 0.010, AP lift = 4.01).

These results effectively bind the direct copious use of trading charts to both technical trading concepts and promotion of paid educational content regarding speculative market activities. This result is particularly fruitful and exemplifies very effectively a successful example of what the declared purpose of this work is; modeling reliable, informative, and most of all usable, relations between visual cue and influencers communication strategies, behavior and intentions.

Another element of novelty stands in the slight relevance of the coin class (0.045). A manual verification assessed that the relevance of this class is mostly due to the object detection model classifying representations of cryptocurrencies as currency, which is coherent with its training process and purpose. This suggests the finfluencers associated with this theme also touch cryptocurrency as sub-topic.

32 - Fintech Infrastructure & Market Access

This theme includes another kind of technical content that does not insist as much on operationally oriented education, but more on information about how financial markets function, such as markets structure, routing orders to financial markets, and novelties about the fintech environment such as stablecoins. The more equilibrated and well-rounded character of this kind of figure, in contrast with finfluencers from the previous case, is well captured by the fact that the model assess similar feature importance for a wide range of them, signaling that trading chart (0.003), suit (0.003) and

tie (0.002) contribute in similar amount to the prediction, while, even if showing low values in absolute terms, coin (0.001) and person (0.001) still give a weak but non-null relative contribution.

Shared and equilibrated feature importance successfully reflect that these finfluencers see rather financial information more as a tool for access to financial markets than speculation, showing that financial dissemination can be consumer- friendly without being predatory. Collectively, these themes demonstrate how finfluencers transform professional finance into modular, consumer-friendly learning platforms, without being too predatory. On the other hand, the weak isolated contribution from the features translates into a noticeable, but far from strong model performance (AUC = 0.656 ± 0.039, AP lift = 1.83).

34 - Promotion of Products, Events, and Community

The finfluencers in this thematic segment have already recognizable characteristics, but eventually indifferent proportions and with a noticeably different focus. While their parasocial relational hooks remain the trading chart (0.029) commentary and speculative instruments such as cryptocurrencies (0.025 as coin), their explicit aim is to create active communities around their professional activity. In doing so they leverage Telegram channels, newsletters, link-in-bio redirection. Engagement prompts, sentiment polls, frequent replies and sometimes live/recorded events represent calls to action that aim to attract customers in a wider community. Issues with this theme is that it presents very few records and that it can easily overlap with *31 - Trader Education & Execution.* While still offering an interesting research opportunity, this greatly compromises its statistical reliability and significance as separate object of inquiry.

35 - Crypto: Between Institutionalization and Hype

The topics found in this theme were bound to emerge as a standalone. Even if they are not novelty anymore, cryptocurrencies are and have been a hot topic for almost the last decade. What is interesting is the framing that is associated with these products. The current discussion about these financial instruments revolves around their progressive institutionalization. On one hand traditional financial products such as ETF already include these assets, some payment methods already integrate them as valid medium, and risk narratives around these instrument have come to be well known. On the other hand, speculative waves drive attention and volatility to these instruments. This ongoing tension is what characterizes the discourse around the topic. Unsurprisingly, finfluencers that talk extensively about cryptocurrencies use their physical symbols very frequently as props and are too easy to recognize through the class coin, but with the effect of other features not being negligible (coin relevance = 0.262, AUC = 0.857 ± 0.018, AP lift = 5.74)

36 - Mindset, Identity, and the Trader Persona

In this last theme another novelty emerges. Finfluencers that constitute this subset cultivate and celebrate their image as presumably successful traders in an overt manner. While a noticeable relevance of classes suit (0.077) and tie (0.065) still suggest concerns still looking professional, and trading chart (0.058) importance still indicates the confrontation with technical topics, these finfluencers shifted their focus on mindset and habits. They still may promote educational programs and try to build a community, but they do so by insisting on a specific narrative. On paper they

promote a journey where patience and discipline are the only discriminant for profitability, but while typically promoting a speculative approach. The potentially pernicious nature of these finfluencers is evident by the fact that the discriminant classes for their identification are for the first time boats (0.090) and cars (0.52); that they use as props.

Even if the contribution of the different features is mostly equilibrated, the peculiarity in the use of props makes associating visual cue to this theme an easy task, since a target that has very salient peculiarities is indeed an easy target, thus allowing for a solid AUC of 0.721 ± 0.012 and an AP lift of 2.81. The use of luxury items as props frames the finfluencers associated with this theme as almost a familiar caricature that is not unique to influencers dedicated to financial topics. The model developed here seems to justify consumer prudency when reached with certain communicational choices.

The results from the random forest analysis also allows for some holistic considerations. For example, not all education is the same. Education concerning budgeting and long-term investing is identified through very different features than education on chart analysis and speculative approaches, remarking the documented distinction between education-first and trading-first accounts (Hammer, 2025). Moreover, the targets that the developed model successfully predicts also loosely map to the five thematic clusters, namely 1) financial literacy basics, budgeting, credit building, 2) diversified long-term investing, retirement planning, 3) active trading, 4) crypto assets and 5) lifestyle change through side income, travel and negotiation scripts, that Pokhrel et al. (2025) documented. Confirmation from recent literature further corroborates the pertinence of the themes

emerged in the thematic analysis and the effectiveness of the random forest model in singling out the relevance of prevalent thematic clusters. Lastly, the choice of the props/object classes that were coded as features, arguably the least validated part of the analysis, demonstrated to be effective; implying that the props that were chosen actually hold semantic and symbolical significance in the context of finfluencers' social media environment.

Proposals for Future Research

The first integration to this study would be to extend the number and nature of props and visual sues to search for in post images and to code into features. Successful addition would improve the performance of the model and maybe allow for effective prediction of themes that were not effectively modeled.

An obvious proposal for expanding the present research is to include additional elements thar are available on Instagram. For example, processing also profile image and profile caption from accounts would allow to create a hierarchical dataset. In the same fashion, also the content of comments would extend the depth of said dataset, but also provide valuable element for assessing the sentiment associated with visual cue and themes.

In addition, presence of visual cues is not the only element that can be extrapolated from post images. Digital creators often include abundant text in images. Optical Character Recognition

(OCR) is a valid technique to extract text from images and integrate it in the analysis, providing way richer material for thematic analysis than only the one found in captions.

Lastly, the most relevant updated to the present work would be implementing a module for monitoring and recurrent collection of content changes in time. That would allow to trace the finfluencers community thematic profile through time and to put the model developed here to the test and to good use. The above suggestions would result in both a hierarchical and longitudinal updating database that would constitute the arrival point of a wider research approach, not only to finfluencers, but to influencers communities in general.

REFERENCES

Abidin, C. (2016). “Aren’t these just young, rich women doing vain things online?”: Influencer selfies as subversive frivolity. Social media+ society, 2(2), 2056305116641342. https://doi.org/10.1177/2056305116641342

Abidin, C. (2018). Internet celebrity: Understanding fame online. Emerald Publishing Limited. https://doi.org/10.1108/978-1-78756-076-520181008

ACM. (2018). ACM Code of Ethics and Professional Conduct. Association for Computing Machinery. https://www.acm.org/code-of-ethics

Ahmed, M., & Othman, R. (2021). PROMOTING OPEN SCIENCE WITH INSTITUTIONAL REPOSITORIES IN THE MALAYSIAN COMPREHENSIVE PUBLIC UNIVERSITIES. Deleted Journal, 3(2), 11-28. https://doi.org/10.31436/jisdt.v3i2.209

Ahmed, S. K., Mohammed, R. A., Nashwan, A. J., Ibrahim, R. H., Abdalla, A. Q., Ameen, B. M. M., & Khidhir, R. M. (2025). Using thematic analysis in qualitative research. Journal of Medicine Surgery and Public Health, 100198. https://doi.org/10.1016/j.glmedi.2025.100198

Alhassan, M. A. M., & Yılmaz, E. (2025). Evaluating YOLOv4 and YOLOv5 for Enhanced Object Detection in UAV-Based Surveillance. Processes, 13(1), 254-269. https://doi.org/10.3390/pr13010254

Ali, M. L., & Zhang, Z. (2024). The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. Computers, 13(12), 336-339. http://dx.doi.org/10.3390/computers13120336

Ali, S. M. S., Dakshinamurthy, T., Priyadarshi, P., & Sanjay, K. (2025). Advances in Consumer research Consumer trust in digital brands: The role of transparency and ethical marketing.

ResearchGate. https://www.researchgate.net/publication/389086224_Advances_in_Consumer_Research_Consumer_Trust_In_Digital_Brands_The_Role_Of_Transparency_And_Ethical_Marketing

Alkentar, S. M., Alsahwa, B., Assalem, A., & Karakolla, D. (2021). Practical comparison of the accuracy and speed of YOLO, SSD, and Faster RCNN for drone detection. Journal of Engineering, 27(8), 19-31. http://dx.doi.org/10.31026/j.eng.2021.08.02

ALLEA. (2019). The European Code of Conduct for Research Integrity -. Allea.org. https://allea.org/code-of-conduct/

Ang, K. G. (2024). The study of the Impact of Financial Goal setting on personal Investment performance. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4844639

Ante, L. (2023). How Elon Musk's twitter activity moves cryptocurrency markets. Technological Forecasting and Social Change, 186, 122112. https://doi.org/10.1016/j.techfore.2022.122112

Anwar, K., Abror, A., Batubara, H. M., Astuti, K., & Sari, N. (2024). Customer engagement and social media research. JPPI (Jurnal Penelitian Pendidikan Indonesia), 10(4), 684-691. https://doi.org/10.29210/020243694

APA. (2017). Ethical principles of psychologists and code of conduct. American Psychological Association. https://www.apa.org/ethics/code

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. Journal of the Academy of Marketing science, 48(1), 79-95. https://doi.org/10.1007/s11747-019-00695-1

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. Science, 365(6456), 858-861. https://doi.org/10.1126/science.aaw8243

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. ACM Transactions on Internet Technology, 1(1), 2-43. https://doi.org/10.1145/383034.383035

Armour, C. (2021). The regulator: Under the finfluence. Company Director, 37(10), 28. https://search.informit.org/doi/10.3316/informit.20220104059595

Aromoye, I. A., Hiung, L. H., & Sebastian, P. (2025). P-DETR: A transformer-based algorithm for pipeline structure detection. Results in Engineering, 26, 104-117. https://doi.org/10.1016/j.rineng.2025.104652

Arora, N., Rana, M., & Prashar, S. (2023). How does social media impact consumers' sustainable purchase intention?. Review of Marketing Science, 21(1), 143-168. https://doi.org/10.1515/roms-2022-0072

Arwidiyarti, D. (2025). Single Shot Multibox Detector (SSD) in Object Detection: A Review. IJACI: International Journal of Advanced Computing and Informatics, 1(2), 118-127. http://dx.doi.org/10.71129/ijaci.v1i2.pp118-127

Ayachi, R., Said, Y., Afif, M., Alshammari, A., Hleili, M., & Abdelali, A. B. (2025). Assessing YOLO models for real-time object detection in urban environments for advanced driver-assistance systems (ADAS). Alexandria Engineering Journal, 123, 530-549. https://doi.org/10.1016/j.aej.2025.03.077

Ayat Abodayeh, Hejazi, R., Najjar, W., Shihadeh, L., & Latif, R. (2023). Web Scraping for Data Analytics: A BeautifulSoup Implementation. https://doi.org/10.1109/wids-psu57071.2023.00025

Aziz, L., Salam, M. S. B. H., Sheikh, U. U., & Ayub, S. (2020). Exploring deep learning-based architecture, strategies, applications, and current trends in generic object detection: A comprehensive review. IEEE Access, 8, 170461-170495. https://doi.org/10.3390/s22062123

Álvarez-Peralta, M., Rojas-Andrés, R., & Diefenbacher, S. (2023). Meta-analysis of political communication research on Twitter: Methodological trends. ProQuest, 9(1). https://doi.org/10.1080/23311886.2023.2209371

Babiarz, P., & Robb, C. A. (2023). Financial literacy and emergency saving. Journal of Family and Economic Issues, 35(1), 40-50. https://doi.org/10.1007/s10834-013-9369-9

Baek, T. H., Kim, J., & Yu, J. H. (2010). The differential roles of brand credibility and brand prestige in consumer brand choice. Psychology & Marketing, 27(7), 662-678. https://doi.org/10.1002/mar.20350

Baghel, D. (2024). Influencer authenticity as a catalyst for brand trust: Analyzing its impact on consumer perception. ShodhKosh Journal of Visual and Performing Arts, 5(6). https://doi.org/10.29121/shodhkosh.v5.i6.2024.3329

Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2022). Attention-induced trading and returns: Evidence from Robinhood users. The Journal of Finance, 77(6), 3141-3190. https://doi.org/10.1111/jofi.13183

Barta, S., Belanche, D., Fernández, A., & Flavián, M. (2023). Influencer marketing on TikTok: The effectiveness of humor and followers' hedonic experience. Journal of Retailing and Consumer Services, 70, 103149. https://doi.org/10.1016/j.jretconser.2022.103149

Beel, J., & Gipp, B. (2010). Academic Search Engine Spam and Google Scholar's Resilience Against It. The Journal of Electronic Publishing, 13(3). https://doi.org/10.3998/3336451.0013.305

Beichert, M., Bayerl, A., Goldenberg, J., & Lanz, A. (2024). Revenue generation through influencer marketing. Journal of Marketing, 88(4), 40-63. https://doi.org/10.1177/00222429231217471

Belanche, D., Casaló, L. V., Flavián, M., & Ibáñez-Sánchez, S. (2021). Building influencers' credibility on Instagram: Effects on followers' attitudes and behavioral responses toward the influencer. Journal of Retailing and Consumer Services, 61, 102585. https://doi.org/10.1016/j.jretconser.2021.102585

Ben-Shmuel, A. T., Hayes, A., & Drach, V. (2024). The gendered language of financial advice: Finfluencers, framing, and subconscious preferences. Socius, 10, 23780231241267131. https://doi.org/10.1177/23780231241267131

Bengani, V. (2025). Humans + LLMs as hybrid teams: Rethinking productivity, creativity, and Decision-Making. ResearchGate. https://doi.org/10.13140/RG.2.2.30355.54560

Best, K. B., Gilligan, J. M., Baroud, H., Carrico, A. R., Donato, K. M., Ackerly, B. A., & Mallick, B. (2021). Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh. Journal of Computational Social Science, 4(1), 77-100. https://doi.org/10.1007/s42001-020-00066-9

Beurer-Kellner, L., Fischer, M., & Vechev, M. (2024). Guiding llms the right way: Fast, non-invasive constrained generation. *arXiv preprint arXiv:2403.06988*.

Beurer-Kellner, L., Fischer, M., & Vechev, M. (2022). Prompting is programming: A query language for large language models. arXiv. https://arxiv.org/abs/2212.06094

Bhatia, R., Bhat, A. K., & Tikoria, J. (2024). Empowering Informed Life insurance Decisions: The impact of Financial literacy on framing effects. Services Marketing Quarterly, 1-29. https://doi.org/10.1080/15332969.2024.2415759

Bischoff, J., Berezan, O., & Scardicchio, L. (2019). The digital self and customer loyalty: from theory to virtual reality. Journal of Marketing Analytics, 7(4), 220-233. https://doi.org/10.1057/s41270-019-00065-4

Björklund, M. (2021). *Beyond moral teaching: Financial literacy as citizenship education* (Doctoral dissertation, Karlstads universitet).

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Bodó, B., Gervais, D., & Quintais, J. P. (2018). Blockchain and Smart contracts: the Missing Link in Copyright licensing? International Journal of Law and Information Technology, 26(4), 311-336. https://doi.org/10.1093/ijlit/eay014

Boerman, S. C., & Van Reijmersdal, E. A. (2020). Disclosing influencer marketing on YouTube to children: The moderating role of para-social relationship. Frontiers in psychology, 10, 3042. https://doi.org/10.3389/fpsyg.2019.03042

Boerman, S. C., Willemsen, L. M., & Van Der Aa, E. P. (2017). "This post is sponsored" effects of sponsorship disclosure on persuasion knowledge and electronic word of mouth in the context of Facebook. Journal of interactive marketing, 38(1), 82-92. https://doi.org/10.1016/j.intmar.2016.12.002

Boesch, G. (2024). GoogLeNet Explained: The Inception Model that Won ImageNet. Viso.ai. https://viso.ai/deep-learning/googlenet-explained-the-inception-model-that-won-imagenet/

Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bornmann, L., Thor, A., Marx, W., & Schier, H. (2016). The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. Journal of the Association for Information Science and Technology, 67(11), 2778-2789. https://doi.org/10.1002/asi.23627

Boston Institute of Analytics. (2025, May 3). The Rise of Finfluencers: How social media is reshaping investment decisions in 2025. Boston Institute of Analytics. https://bostoninstituteofanalytics.org/blog/the-rise-of-finfluencers-how-social-media-is-reshaping-investment-decisions-in-2025/

Boumans, J. W., & Trilling, D. (2016). Taking Stock of the Toolkit. Digital Journalism, 4(1), 8-23. https://doi.org/10.1080/21670811.2015.1096598

Bouraya, S., & Belangour, A. (2021). Object detectors: convolutional neural networks backbones: A review and a comparative study. International Journal, 9(11), 1379-1385. http://dx.doi.org/10.30534/ijeter/2021/039112021

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2022). Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. International Journal of Transgender Health, 24(1), 1-6. https://doi.org/10.1080/26895269.2022.2129597

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), 107-117. https://doi.org/10.1016/s0169-7552(98)00110-x

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Brokensha, S., Kotzé, E., & Senekal, B. A. (2019). Reinventing the social scientist and humanist in the era of big data - A perspective from South African scholars. *SunBonani Scholar*. https://doi.org/10.18820/9781928424376

Bruns, A., & Burgess, J. (2011). The Use of Twitter Hashtags in the Formation of Ad Hoc Publics. https://eprints.qut.edu.au/46515/1/The_Use_of_Twitter_Hashtags_in_the_Formation_of_Ad_Hoc_Publics_(final).pdf

Burgess, M. (2025). How advisors can respond to the finfluencer challenge. Globe and Mail web edition. https://link.gale.com/apps/doc/A837046698/AONE?u=anon~4dfde446&sid=googleScholar&xid=8371c896

Burns, M., Bally, J., Burles, M., Holtslander, L., & Peacock, S. (2022). Constructivist grounded theory or interpretive phenomenology? Methodological choices within specific study

contexts. International Journal of Qualitative Methods, 21. https://doi.org/10.1177/16094069221077758

Cadwalladr, C., & Graham-Harrison, E. (2018, March 18). Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach. The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Campbell, C., & Grimm, P. E. (2019). The challenges native advertising poses: Exploring potential federal trade commission responses and identifying research needs. Journal of Public Policy & Marketing, 38(1), 110-123. https://doi.org/10.1177/0743915618818576

Campbell, M. C., Inman, J. J., Kirmani, A., & Price, L. L. (2020). In times of trouble: A framework for understanding consumers' responses to threats. Journal of consumer research, 47(3), 311-326. https://doi.org/10.1093/jcr/ucaa036

Canatan, E. C., Toker, A., & Coşkun, A. (2023). Understanding Finfluencer Engagement: A Conceptual Framework of Attitude Development and Continued Usage in Video Consumption. https://aisel.aisnet.org/menacis2023/17

Cannata, M., Collombin, M., Ertz, O., Giuliani, G., Ingensand, J., Primerano, C., & Strigaro, D. (2023). *The Challenges of Reproducibility for Research Based on Geodata Web Services*. https://doi.org/10.20944/preprints202312.2316.v1

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with Transformers. Lecture Notes in Computer Science, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)* (pp. 2633-2650).

Casaló, L. V., Flavián, C., & Ibáñez-Sánchez, S. (2020). Influencers on Instagram: Antecedents and consequences of opinion leadership. Journal of business research, 117, 510-519. https://doi.org/10.1016/j.jbusres.2018.07.005

Chadwick, A., Vaccari, C. & O'Loughlin, B. (2018) Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. New Media & Society, 20 (11), 4255-4274. doi:10.1177/1461444818769689

Chahal, K. S., & Dey, K. (2018). A survey of modern object detection literature using deep learning. arXiv preprint arXiv:1808.07256. https://arxiv.org/pdf/1808.07256

Chang, C. Y., & He, X. (2025). The Liabilities of Robots. https://doi.org/10.2139/ssrn.5159436

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Health and Surveillance, 6(2), e19273. https://doi.org/10.2196/19273

Cheng, K., Li, Z., Sun, Z., Guo, Q., Li, W., Lu, Y., Qi, S., Shen, Z., Xie, R., Wang, Y., Wu, Z., Wu, Y., Wu, C., Li, Y., Xie, Y., Wu, H., & Li, C. (2024). The rapid growth of bibliometric studies: a call for international guidelines. International Journal of Surgery, 110(4), 2446-2448. https://doi.org/10.1097/js9.0000000000001049

Chiarello, F., Giordano, V., Spada, I., Barandoni, S., & Fantoni, G. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. Technovation, 133, 103002. https://doi.org/10.1016/j.technovation.2024.103002

Cho, J., & Garcia-Molina, H. (2002). Parallel crawlers. 1-13. https://doi.org/10.1145/511446.511464

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Christou, P. A. (2022). How to use thematic analysis in qualitative research. Journal of Qualitative Research, 3(2), 79-95. https://doi.org/10.4337/jqrt.2023.0006

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. Scientific reports, 10(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Cong, X., Li, S., Chen, F., Liu, C., & Meng, Y. (2023). A review of YOLO object detection algorithms based on deep learning. Frontiers in Computing and Intelligent Systems, 4(2), 17-20. http://dx.doi.org/10.54097/fcis.v4i2.9730

Cookson, J. A., Fox, C., Gil-Bazo, J., Imbet, J. F., & Schiller, C. (2023). Social media as a bank run catalyst. Available at SSRN, 4422754. http://dx.doi.org/10.2139/ssrn.4422754

Cornwell, T. B., & Kwon, Y. (2020). Sponsorship-linked marketing: Research surpluses and shortages. Journal of the Academy of Marketing Science, 48(4), 607-629. https://doi.org/10.1007/s11747-019-00654-w

Costa, M., Gomes, D., & Silva, M. J. (2016). The evolution of web archiving. International Journal on Digital Libraries, 18(3), 191-205. https://doi.org/10.1007/s00799-016-0171-9

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. New media & society, 21(4), 895-913. https://doi.org/10.1177/1461444818815684

Crasta, S. (2024). ENHANCING CONSUMER VIGILANCE AND MITIGATING TACTICS AGAINST INTERNET SHOPPING FRAUD. *Al-Shodhana*, *12*(2), 86-93.

Cuello, C. (2024, December 14). *Data Extraction: Techniques, tools, and real-time benefits*. Rivery. https://rivery.io/data-learning-center/data-extraction/

De Castro, C. A. (2023). Thematic analysis in social media influencers: who are they following and why? Frontiers in Communication, 8. https://doi.org/10.3389/fcomm.2023.1217684

De Jans, S., Cauberghe, V., & Hudders, L. (2018). How an advertising disclosure alerts young adolescents to sponsored vlogs: The moderating role of a peer-based advertising literacy intervention through an informational vlog. Journal of Advertising, 47(4), 309-325. https://doi.org/10.1080/00913367.2018.1539363

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, 3(1). https://doi.org/10.1038/srep01376

De Paoli, S., & Mathis, W. S. (2024). Reflections on inductive thematic saturation as a potential metric for measuring the validity of an inductive thematic analysis with LLMs. Quality & Quantity. https://doi.org/10.1007/s11135-024-01950-6

De Veirman, M., & Hudders, L. (2020). Disclosing sponsored Instagram posts: the role of material connection with the brand and message-sidedness when disclosing covert advertising. International journal of advertising, 39(1), 94-130. https://doi.org/10.1080/02650487.2019.1575108

De Veirman, M., Cauberghe, V., & Hudders, L. (2017). Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude. International journal of advertising, 36(5), 798-828. https://doi.org/10.1080/02650487.2017.1348035

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, *35*, 30318-30332.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems (NeurIPS 2025). https://neurips.cc/media/neurips-2023/Slides/73855.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL. https://aclanthology.org/N19-1423/

Djafarova, E., & Bowes, T. (2021). 'Instagram made Me buy it': Generation Z impulse purchases in fashion industry. Journal of retailing and consumer services, 59, 102345. https://doi.org/10.1016/j.jretconser.2020.102345

Djafarova, E., & Rushworth, C. (2017). Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users. Computers in human behavior, 68, 1-7. https://doi.org/10.1016/j.chb.2016.11.009

Dogucu, M., & Çetinkaya-Rundel, M. (2020). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. Journal of Statistics Education, 29(1), 1-11. https://doi.org/10.1080/10691898.2020.1787116

Dubois, D., Bonezzi, A., & De Angelis, M. (2016). Sharing with friends versus strangers: How interpersonal closeness influences word-of-mouth valence. Journal of Marketing Research, 53(5), 712-727. https://doi.org/10.1509/jmr.13.0312

Duka, M., Sikora, M., & Strzelecki, A. (2023). From Web Catalogs to Google: A Retrospective Study of Web Search Engines Sustainable Development. Sustainability, 15(8), 6768. https://doi.org/10.3390/su15086768

Durve, M., Orsini, S., Tiribocchi, A., Montessori, A., Tucny, J.-M., Lauricella, M., Camposeo, A., Pisignano, D., & Succi, S. (2023). Benchmarking Yolov5 and Yolov7 models with DeepSORT for droplet tracking applications. *The European Physical Journal E*, *46*(5). https://doi.org/10.1140/epje/s10189-023-00290-x

Dwork, C. (2008). Differential Privacy: A Survey of Results. Lecture Notes in Computer Science, 4978, 1-19. https://doi.org/10.1007/978-3-540-79228-4_1

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational Social Science and Sociology. Annual Review of Sociology, 46(1), 61-81. https://doi.org/10.1146/annurev-soc-121919-054621

Edozie, E., Shuaibu, A. N., John, U. K., & Sadiq, B. O. (2025). Comprehensive review of recent developments in visual object detection based on deep learning. Artificial Intelligence Review, 58(9), 277-289. https://doi.org/10.1007/s10462-025-11284-w

Erkan, I., & Evans, C. (2018). Social media or shopping websites? The influence of eWOM on consumers' online purchase intentions. Journal of marketing communications, 24(6), 617-632. https://doi.org/10.1080/13527266.2016.1184706

Eken, B., Pallewatta, S., Tran, N., Tosun, A., & Babar, M. A. (2025). A multivocal review of Mlops Practices, challenges and open issues. *ACM Computing Surveys*, *58*(2), 1–35. https://doi.org/10.1145/3747346

El-Moussaoui, M., Hanine, M., Kartit, A., Villar, M. G., Garay, H., & de la Torre Díez, I. (2025). A systematic review of deep learning methods for community detection in social networks. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1572645

European Securities and Markets Authority (ESMA). (2024). Annual report 2023 (ESMA22-50751485-1453). Publications Office of the European Union. https://data.europa.eu/doi/10.2856/28329

Evans, N. J., Phua, J., Lim, J., & Jun, H. (2017). Disclosing Instagram influencer advertising:

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338. http://dx.doi.org/10.1007/s11263-009-0275-4

Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. BMJ, 323(7321), 1103-1105. https://doi.org/10.1136/bmj.323.7321.1103

Fairfield, J., & Engel, C. (2015). PRIVACY AS A PUBLIC GOOD. Duke Law Journal, 65(3), 1-73. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3824&context=dlj

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. arXiv. https://arxiv.org/abs/2101.03961

Feng, Y., You, Y., Tian, J., & Meng, G. (2023). OEGR-DETR: A novel detection transformer based on orientation enhancement and group relations for SAR object detection. Remote Sensing, 16(1), 106-118. https://doi.org/10.3390/rs16010106

Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, *70*, 301–323. https://doi.org/10.1016/j.knosys.2014.07.007

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. Social Media + Society, 4(1). https://doi.org/10.1177/2056305118763366

Fife, D. A., & D'Onofrio, J. (2023). Common, uncommon, and novel applications of random forest in psychological research. Behavior Research Methods, 55(5), 2447-2466. https://doi.org/10.3758/s13428-022-01901-9

Fife, S. T., & Gossner, J. D. (2024). Deductive Qualitative analysis: evaluating, expanding, and refining theory. International Journal of Qualitative Methods, 23. https://doi.org/10.1177/16094069241244856

Flaherty, G. T., & Mangan, R. M. (2025). Impact of social media influencers on amplifying positive public health messages. Journal of Medical Internet Research, 27, e73062. https://doi.org/10.2196/73062

Floridi, L., & Taddeo, M. (2016). What Is Data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083). https://doi.org/10.1098/rsta.2016.0360

Foerster, S., Linnainmaa, J. T., Melzer, B. T., & Previtero, A. (2017). Retail financial advice: does one size fit all?. The Journal of Finance, 72(4), 1441-1482. https://doi.org/10.1111/jofi.12514

Fornero, E., & Lo Prete, A. (2023). Financial education: From better personal finance to improved citizenship. Journal of Financial Literacy and Wellbeing, 1(1), 12-27. https://doi.org/10.1017/flw.2023.7

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Galvao, L. G., Abbod, M., Kalganova, T., Palade, V., & Huda, M. N. (2021). Pedestrian and vehicle detection in autonomous vehicle perception systems - A review. Sensors, 21(21), 7267-7287. https://doi.org/10.3390/s21217267

Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y., & Zhu, Y. (2024). A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. *Expert Systems with Applications*, *242*, 122731. https://doi.org/10.1016/j.eswa.2023.122731

Gao, P., Zheng, M., Wang, X., Dai, J., & Li, H. (2021). Fast convergence of detr with spatially modulated co-attention. In Proceedings of the IEEE/CVF international conference on computer vision. 3621-3630. http://dx.doi.org/10.1109/ICCV48922.2021.00360

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6894-6910). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.552

Gass, R. H., & Seiter, J. S. (2022). Persuasion: Social influence and compliance gaining. Routledge. https://doi.org/10.4324/9781003081388

GDPR. (n.d.). General Data Protection Regulation (GDPR). GDPR. https://gdpr-info.eu/

Geiger, C., Frosio, G., & Bulayenko, O. (2018). The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects | Think Tank | European Parliament. Europa.eu. https://www.europarl.europa.eu/thinktank/en/document/IPOL_IDA(2018)604941

Gerritsen, D., & de Regt, A. (2025). Influencers and Consumer Financial Decision-Making. International Journal of Consumer Studies, 49(2), e70037. https://doi.org/10.1111/ijcs.70037

Ghadafi, E., & Andriotis, P. (2025, May 3). UK finfluencers: Exploring content, reach, and responsibility. arXiv. https://arxiv.org/html/2505.01941v1

Ghanaei, Z., & Rouhani, M. (2025). A context-aware multiclass loss function for semantic segmentation with a focus on intricate areas and class imbalances. Scientific Reports, 15(1). 1-17. https://doi.org/10.1038/s41598-025-08234-5

Gillings, M., Kohn, T., & Mautner, G. (2024). The rise of large language models: challenges for Critical Discourse Studies. Critical Discourse Studies, 1-17. https://doi.org/10.1080/17405904.2024.2373733

Girshick, R. J. C. S. (2015). Fast R-CNN. arXiv 2015. arXiv preprint arXiv:1504.08083, 729. https://doi.org/10.48550/arXiv.1504.08083

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 580-587. https://doi.org/10.1109/CVPR.2014.81

Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., & Nida-Rümelin, J. (2021). Ethics in the Software Development Process: from Codes of Conduct to Ethical Deliberation. Philosophy & Technology, 34(4). https://link.springer.com/article/10.1007/s13347-021-00451-w

Gomber, P., Koch, J. A., & Siering, M. (2017). Digital Finance and FinTech: current research and future research directions. Journal of business economics, 87(5), 537-580. https://doi.org/10.1007/s11573-017-0852-x

Google. (2024). Gemma 2 is now available to researchers and developers. Google AI Blog. https://blog.google/technology/developers/google-gemma-2/

Goud, N. K. A., Kumar, N. D. K. V. R. S., & PChakradhar, N. D. (2024). A study on behavioural finance and its impact on decision making of an investment. EPRA International Journal of Economics, Business and Management Studies, 11(1), 104-115. https://doi.org/10.36713/epra16186

Govindarajan, V., Srivastava, A., & Chatterjee, C. (2025, January 24). How "finfluencers" can create risk for your company. Harvard Business Review. https://hbr.org/2025/01/how-finfluencers-can-create-risk-for-your-company

Graf-Vlachy, L., Buhtz, K., & König, A. (2018). Social influence in technology adoption: taking stock and moving forward. Management Review Quarterly, 68(1), 37-76. https://doi.org/10.1007/s11301-017-0133-3

Green, D. D., Walker, C., Alabulththim, A., Smith, D., & Phillips, M. (2018). Fueling the Gig Economy: A Case Study Evaluation of Upwork.com. Management and Economics Research Journal, 04(2018), 104-112. https://doi.org/10.18639/merj.2018.04.523634

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science advances, 5(1), eaau4586. DOI:10.1126/sciadv.aau4586

Guetterman, T. C., Fetters, M. D., & Creswell, J. W. (2015). Integrating quantitative and qualitative results in health science mixed methods research through joint displays. The Annals of Family Medicine, 13(6), 554-561. https://doi.org/10.1370/afm.1865

Gulli, A., & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. Special Interest Tracks and Posters of the 14th International Conference on World Wide Web - WWW '05, 902. https://doi.org/10.1145/1062745.1062789

Guo, Q., Jiao, S., Yang, Y., Yu, Y., & Pan, Y. (2025). Assessment of urban flood disaster responses and causal analysis at different temporal scales based on social media data and machine learning algorithms. *International Journal of Disaster Risk Reduction*, *117*, 105170. https://doi.org/10.1016/j.ijdrr.2024.105170

Gutierrez, G., Llerena, J. P., Usero, L., & Patricio, M. A. (2024). A comparative study of convolutional neural network and transformer architectures for drone detection in thermal images. Applied Sciences, 15(1), 109-118. http://dx.doi.org/10.3390/app15010109

Guyt, J. Y., Datta, H., & Boegershausen, J. (2024). Unlocking the Potential of Web Data for Retailing Research. Journal of Retailing, 100(1). https://doi.org/10.1016/j.jretai.2024.02.002

Haase, F., Rath, O., Krauß, J., & Schoder, D. (2025). The role of finfluencers in shaping crowd sentiment. Business & Information Systems Engineering, 67(2), 115-132. https://doi.org/10.1007/s12599-025-00947-1

Haase, F., Rath, O., Kurka, M., & Schoder, D. (2023). Finfluencers: Opinion makers or opinion followers?. https://aisel.aisnet.org/ecis2023_rp/432

Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. International Journal of Intelligent Networks, 3, 119-132. http://dx.doi.org/10.1016/j.ijin.2022.08.005

Hammer, C. C. (2025). The Role of Social Media in Shaping Investment Trends Among College Students. Finance Undergraduate Honors Theses Retrieved from https://scholarworks.uark.edu/finnuht/149

Harmeling, C. M., Moffett, J. W., Arnold, M. J., & Carlson, B. D. (2017). Toward a theory of customer engagement marketing. Journal of the Academy of marketing science, 45(3), 312-335. https://doi.org/10.1007/s11747-016-0509-2

Hasan, M. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. Journal of Big Data, 7(1), 21. https://doi.org/10.1186/s40537-020-00291-z

Hasanah, E. N., Koesrindartoto, D. P., Wiryono, S. K., & Angelica, A. E. (2025). Who deserves to be the finfluencer? Journal of Open Innovation Technology Market and Complexity, 100553. https://doi.org/10.1016/j.joitmc.2025.100553

Hayes, A. S. (2025). "Conversing" with Qualitative data: Enhancing qualitative research through large language models (LLMs). International Journal of Qualitative Methods, 24. https://doi.org/10.1177/16094069251322346

Hayes, A. S., & Ben-Shmuel, A. T. (2024). Under the finfluence: Financial influencers, economic meaning-making and the financialization of digital life. Economy and Society, 53(3), 478-503. https://doi.org/10.1080/03085147.2024.2381980

He, H., Xu, H., Zhang, Y., Gao, K., Li, H., Ma, L., & Li, J. (2022). Mask R-CNN-based automated identification and extraction of oil well sites. International Journal of Applied Earth Observation and Geoinformation, 112, 1-12. https://doi.org/10.1016/j.jag.2022.102875

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

He, L. H., Zhou, Y. Z., Liu, L., Cao, W., & Ma, J. H. (2025). Research on object detection and recognition in remote sensing images based on YOLOv11. Scientific Reports, 15(1), 1-14. http://dx.doi.org/10.1038/s41598-025-96314-x

Henderson, P., & Ferrari, V. (2016). End-to-end training of object class detectors for mean average precision. In the Asian Conference on Computer Vision (pp. 198-213). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-54193-8_13

Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible Web crawler. World Wide Web, 2(4), 219-229. https://doi.org/10.1023/a:1019213109274

Higgins, J. P. T. & Green, S. (2023). Cochrane handbook for systematic reviews of interventions (v.6.4). Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook

Hii, I. S., & Ong, Y. X. (2025). Finfluencer: can financial social media influencers promote desirable financial behaviours?. International Journal of Bank Marketing, 1-29. https://doi.org/10.1108/IJBM-05-2024-0256

Hitchcock, T. (2023, June 6). Under the influence: Regulatory responses to financial promotions by social media influencers. Thomson Reuters Institute. https://www.thomsonreuters.com/en-us/posts/investigation-fraud-and-risk/finfluencers-regulatory-response/

Hoffman, D. L., & Novak, T. P. (2018). Consumer and object experience in the internet of things: An assemblage theory approach. Journal of Consumer Research, 44(6), 1178-1204. https://doi.org/10.1093/jcr/ucx105

Holloway, S. L., & Pimlott-Wilson, H. (2021). Solo self-employment, entrepreneurial subjectivity and the security-precarity continuum: Evidence from private tutors in the supplementary education industry. Environment and Planning a Economy and Space, 53(6), 1547-1564. https://doi.org/10.1177/0308518x211009237

Hosseini, A., Hooshanfar, K., Omrani, P., Toosi, R., Toosi, R., Ebrahimian, Z., & Akhaee, M. A. (2025). Brand visibility in packaging: A deep learning approach for logo detection, saliency-map prediction, and logo placement analysis. Discover Applied Sciences, 7(6), 537-548. https://doi.org/10.1007/s42452-025-07043-9

Hridoy, Md. T., Saha, S. R., Islam, M. M., Uddin, M. A., & Mahmud, Md. Z. (2024). Leveraging web scraping and stacking ensemble machine learning techniques to enhance detection of major depressive disorder from social media posts. *Social Network Analysis and Mining*, *14*(1). https://doi.org/10.1007/s13278-024-01392-w

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Hu, R., Zhang, D., Tao, D., Hartvigsen, T., Feng, H., & Rundensteiner, E. (2022, September 14). *Tweet-FID: An annotated dataset for multiple foodborne illness detection tasks*. arXiv.org. https://doi.org/10.48550/arXiv.2205.10726

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L. J., Xiao, G., Peterson, E. D., Zhan, X., & Xie, Y. (2024). A critical assessment of

using CHATGPT for extracting structured data from clinical notes. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01079-8

Huang, S., Aral, S., Hu, Y. J., & Brynjolfsson, E. (2020). Social advertising effectiveness across products: A large-scale field experiment. Marketing Science, 39(6), 1142-1165. https://doi.org/10.1287/mksc.2020.1240

Hudders, L., De Jans, S., & De Veirman, M. (2021). The commercialization of social media stars. International Journal of Advertising, 40(3), 327-375. https://doi.org/10.1080/02650487.2020.1836925

Hudson, S., Roth, M. S., Madden, T. J., & Hudson, R. (2015). The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees. Tourism management, 47, 68-76. https://doi.org/10.1016/j.tourman.2014.09.001

Hugging Face. (2025). Archived Open LLM Leaderboard (2024-2025). https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

Hugh Wilkie, D. C., Dolan, R., Harrigan, P., & Gray, H. (2022). Influencer marketing effectiveness: the mechanisms that matter. European Journal of Marketing, 56(12), 3485-3515. https://doi.org/10.1108/EJM-09-2020-0703

Hull, I., & Qi, Y. (2024). The impact of finfluencers on retail investment. Available at SSRN 4922031. http://dx.doi.org/10.2139/ssrn.4922031

Hussain, S., Mubeen, I., Ullah, N., Shah, S. S. U. D., Khan, B. A., Zahoor, M., ... & Sultan, M. A. (2022). Modern diagnostic imaging technique applications and risk factors in the medical field: a review. BioMed Research International, 2022(1), 1-16. https://doi.org/10.1155/2022/5164970

Hwang, K., & Zhang, Q. (2018). Influence of parasocial relationship between digital celebrities and their followers on followers' purchase and electronic word-of-mouth intentions, and persuasion knowledge. Computers in human behavior, 87, 155-173. https://doi.org/10.1016/j.chb.2018.05.029

IAPP. (2024). The state of web scraping in the EU. Iapp.org. https://iapp.org/news/a/the-state-of-web-scraping-in-the-eu

IBM Technology. (2025, April 8). What is Ollama? Running Local LLMs Made Simple [Video]. YouTube. https://www.youtube.com/watch?v=5RIOQuHOihY

ILO. (2024, July 9). Pay transparency can address the gender pay gap. International Labour Organization. https://www.ilo.org/resource/news/pay-transparency-can-address-gender-pay-gap

Issac, A., & Seranmadevi, R. (2024). Revolutionizing financial literacy through exploring finfluencers' intentions in Chatbot-Driven financial information dissemination. In Advances in business information systems and analytics book series (pp. 159-188). https://doi.org/10.4018/979-8-3693-4187-2.ch008

Jain, S., van Zuylen, M., Hajishirzi, H., & Beltagy, I. (2020). SciREX: Document-level information extraction for scientific articles. ACL. https://aclanthology.org/2020.acl-main.670/

Jamali, M., Davidsson, P., Khoshkangini, R., Ljungqvist, M. G., & Mihailescu, R. C. (2025). Context in object detection: a systematic literature review. *Artificial Intelligence Review*, *58*(6), 1-89.

Jantan, M. S. (2024). Analysis on Sales Appropriateness for Finfluencer. Available at SSRN 4803657. http://dx.doi.org/10.2139/ssrn.4803657

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis - WebKDD/SNA-KDD '07. https://doi.org/10.1145/1348549.1348556

Javare, P., Khetan, D., Kamerkar, C., Gupte, Y., Chachra, S., & Joshi, U. (2020, April). Using object detection and data analysis for developing customer insights in a retail setting. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).

Jensen, M., Brix Danielsen, M., Riis, J., Assifuah Kristjansen, K., Andersen, S., Okubo, Y., & Jørgensen, M. G. (2025). CHATGPT-4O can serve as the second rater for data extraction in systematic reviews. *PLOS ONE*, *20*(1). https://doi.org/10.1371/journal.pone.0313401

Jegham, N., Koh, C. Y., Abdelatti, M., & Hendawi, A. (2024). Evaluating the evolution of yolo (You Only Look Once) models: A comprehensive benchmark study of YOLOv11 and its predecessors. arXiv e-prints, arXiv-2411. https://ui.adsabs.harvard.edu/link_gateway/2024arXiv241100201J/doi:10.48550/arXiv.2411.00201

Jiang, D., Liu, Y., Liu, S., Zhao, J. E., Zhang, H., Gao, Z., ... & Xiong, H. (2023). From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Jin, S. V., Muqaddam, A., & Ryu, E. (2019). Instafamous and social media influencer marketing. Marketing intelligence & planning, 37(5), 567-579. https://doi.org/10.1108/MIP-09-2018-0375

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs (FAISS). arXiv. https://arxiv.org/abs/1702.08734

Johnstone, L., & Lindh, C. (2018). The sustainability-age dilemma: A theory of (un) planned behaviour via influencers. Journal of consumer behaviour, 17(1), e127-e139. https://doi.org/10.1002/cb.1693

Jones, Z., & Linder, F. (2015). Exploratory data analysis using random forests. In Prepared for the 73rd annual MPSA conference (pp. 1-31).

Jungherr, A. (2015). Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research. Springer International Publishing.

Jungherr, A., Schoen, H., & Jürgens, P. (2015). The Mediation of Politics through Twitter: An Analysis of Messages posted during the Campaign for the German Federal Election 2013. Journal of Computer-Mediated Communication, 21(1), 50-68. https://doi.org/10.1111/jcc4.12143

Justia Law. (2013). The Associated Press V. Meltwater US Holdings, Inc. et al, No. 1:2012cv01087 - Document 156 (S.D.N.Y. 2013). Justia Law. https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2012cv01087/392003/156/

Justia Law. (2022). HIQ LABS, INC. V. LINKEDIN CORPORATION, No. 17-16783 (9th Cir. 2022). Justia Law. https://law.justia.com/cases/federal/appellate-courts/ca9/17-16783/17-16783-2022-04-18.html

Kalake, L., Dong, Y., Wan, W., & Hou, L. (2022). Enhancing detection quality rate with a combined hog and CNN for real-time multiple object tracking across non-overlapping multiple cameras. Sensors, 22(6), 1-20. https://doi.org/10.3390/s22062123

Kamal, S. A., Mohammed, A. Razak, M. Z., Abd Rahman, A. H., & Bakar, M. A. (2024). An efficient intersection over Union algorithm for 3D object detection. IEEE Access.

Kang, S., Hu, Z., Liu, L., Zhang, K., & Cao, Z. (2025). Object detection YOLO algorithms and their industrial applications: Overview and comparative analysis. Electronics, 14(6), 1-36. https://doi.org/10.3390/electronics14061104

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6769-6781). Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.550/

Karwatzki, S., Dytynko, O., Trenz, M., & Veit, D. (2017). Beyond the Personalization-Privacy Paradox: Privacy Valuation, Transparency Features, and Service Personalization. Journal of Management Information Systems, 34(2), 369-400. https://doi.org/10.1080/07421222.2017.1334467

Kay, S., Mulcahy, R., & Parkinson, J. (2020). When less is more: the impact of macro and micro social media influencers' disclosure. Journal of marketing management, 36(3-4), 248-278. https://doi.org/10.1080/0267257X.2020.1718740

Kedvarin, S., & Saengchote, K. (2023). Social media finfluencers: evidence from YouTube and cryptocurrencies. Available at SSRN 4594081. http://dx.doi.org/10.2139/ssrn.4594081

Khalfallah, D., & Keller, V. (2025). Authenticity, ethics, and transparency in virtual influencer marketing: A cross-cultural analysis of consumer trust and engagement: A systematic literature review. Acta Psychologica, 260, 105573. https://doi.org/10.1016/j.actpsy.2025.105573

Khan, F., Siddiqui, M. A., & Imtiaz, S. (2022). Role of financial literacy in achieving financial inclusion: A review, synthesis and research agenda. Cogent Business & Management, 9(1). https://doi.org/10.1080/23311975.2022.2034236

Khan, M. S., & Imran, A. (2024). The Art of Seeing: A Computer Vision Journey into Object Detection. Adv Mach Lear Art Inte, 5(2), 1-6. http://dx.doi.org/10.21203/rs.3.rs-4361138/v1

Khandolkar, T., Desai, P. H., Mekoth, N., & Borde, N. (2024). Social media brand engagement and perceived risk in purchase: a conceptual framework. ReMark-Revista Brasileira de Marketing, 23(4), 1858-1883. https://doi.org/10.5585/remark.v23i4.25612

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., ... & Potts, C. (2023). Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Khder, M. A. (2021). Web Scraping or Web Crawling: State of the Art, Techniques, Approaches, and Applications. International Journal of Advances in Soft Computing and Its Applications, 13(3), 145-168. https://doi.org/10.15849/ijasca.211128.11

Khoirotunnisa, F. (2024). From Advice to Action: How Finfluencers are Reshaping Investment Behavior. Journal of Economics, Business, and Government Challenges, 7(01), 48-57. https://doi.org/10.33005/ebgc.v7i01.1530

Khurana, K. (2023). Finfluencers as investment advisors-Time to rein them in?. University of Michigan. https://dx.doi.org/10.7302/7938

Ki, C. W. C., Cuevas, L. M., Chong, S. M., & Lim, H. (2020). Influencer marketing: Social media influencers as human brands attaching to followers and yielding positive marketing

results by fulfilling needs. Journal of retailing and consumer services, 55, 102133. https://doi.org/10.1016/j.jretconser.2020.102133

Kicova, E., Michulek, J., Ponisciakova, O., & Fabus, J. (2025). When financial awareness meets reality: financial literacy and Gen Z's entrepreneurship interest. International Journal of Financial Studies, 13(3), 171. https://doi.org/10.3390/ijfs13030171

Kim, M., & Kim, J. (2021). Corporate Social Responsibility, Employee Engagement, Well-Being and the Task Performance of Frontline Employees. Management Decision, 59, 2040-2056. https://doi.org/10.1108/MD-03-2020-0268

Kim, D. Y., & Kim, H. Y. (2023). Social media influencers as human brands: an interactive marketing perspective. *Journal of Research in Interactive Marketing*, *17*(1), 94-109.

Kim, S. J., Wang, R. J. H., & Malthouse, E. C. (2015). The effects of adopting and using a brand's mobile application on customers' subsequent purchase behavior. Journal of Interactive Marketing, 31(1), 28-41. https://doi.org/10.1016/j.intmar.2015.05.004

Kishor, R. (2024). Performance Benchmarking of YOLOv11 Variants for Real-Time Delivery Vehicle Detection: A Study on Accuracy, Speed, and Computational Trade-offs. Asian Journal of Research in Computer Science, 17(12), 108-122. http://dx.doi.org/10.9734/ajrcos/2024/v17i12532

Kizgin, H., Jamal, A., Dey, B. L., & Rana, N. P. (2018). The impact of social media on consumers' acculturation and purchase intentions. Information Systems Frontiers, 20(3), 503-514. https://doi.org/10.1007/s10796-017-9817-4

Knura, M., Kluger, F., Zahtila, M., Schiewe, J., Rosenhahn, B., & Burghardt, D. (2021). Using object detection on social media images for Urban Bicycle Infrastructure Planning: A case

study of Dresden. *ISPRS International Journal of Geo-Information*, *10*(11), 733. https://doi.org/10.3390/ijgi10110733

Kolobov, A., Peres, Y., Lubetzky, E., & Horvitz, E. (2019). Optimal Freshness Crawl Under Politeness Constraints. 495-504. https://doi.org/10.1145/3331184.3331241

Krause, D. (2025). The Impact of Financial Influencers on Crypto Markets: Systemic Risks and Regulatory Challenges. Available at SSRN 5144847. http://dx.doi.org/10.2139/ssrn.5144847

Krämer, N. C., Winter, S., Benninghoff, B., & Gallus, C. (2015). How "social" is Social TV? The influence of social motives and expected outcomes on the usage of Social TV applications. Computers in human behavior, 51, 255-262. https://doi.org/10.1016/j.chb.2015.05.005

Krichen, M. (2023). Convolutional neural networks: A survey. Computers, 12(8), 151-163. https://doi.org/10.3390/computers12080151

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. https://doi.org/10.1145/3065386

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

Kumar, C. (2025). Finance influencers: A literature review paper. ResearchGate. https://www.researchgate.net/publication/395808459_Finance_Influencers_A_Literature_Review_Paper

Kumar, R., Jain, A., & Agrawal, Chetan. (2014). Survey of Web Crawling Algorithms. SSRN Electronic Journal, 1(2/3). https://doi.org/10.2139/ssrn.3437184

Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. ACM SIGMOD Record, 31(2), 84. https://doi.org/10.1145/565117.565137

Lai, K. P. (2025). FinTech: making finance fun. In A Research Agenda for Economic Geography (pp. 175-188). Edward Elgar Publishing. https://doi.org/10.4337/9781035339921.00019

Lalwani, V. (2025). Finfluencer recommendations. Economics Letters, 112511. https://doi.org/10.1016/j.econlet.2025.112511

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Leung, F. F., Gu, F. F., & Palmatier, R. W. (2022). Online influencer marketing. Journal of the Academy of Marketing Science, 50(2), 226-251. https://doi.org/10.1007/s11747-021-00829-4

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

Li, H., & Zhang, N. (2024). Computer Vision Models for Image Analysis in Advertising Research. Journal of Advertising, 53(5), 771-790. http://dx.doi.org/10.1080/00913367.2024.2407644

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in Natural Language Processing: A Survey. *AI Open*, *3*, 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001

Li, J., Li, D., Xiong, C., & Hoi, S. (2022, February 15). *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*. arXiv.org. https://doi.org/10.48550/arXiv.2201.12086

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., & Sun, J. (2018). Detnet: A backbone network for object detection. arXiv preprint arXiv:1804.06215. https://doi.org/10.48550/arXiv.1804.06215

Lillywhite, B., & Wolbring, G. (2022). Emergency and Disaster Management, Preparedness, and Planning (EDMPP) and the 'Social': A scoping review. Sustainability, 14(20), 13519. https://doi.org/10.3390/su142013519

Lin, H., & Chen, Q. (2024). Artificial intelligence (AI)-integrated educational applications and college students' creativity and academic emotions: students and teachers' perceptions and attitudes. BMC Psychology, 12(1), 487-493. https://doi.org/10.1186/s40359-024-01979-0

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. ACL. https://aclanthology.org/2022.acl-long.229/

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Cham: Springer International Publishing.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single-shot multibox detector. In the European conference on computer vision. 21-37. https://doi.org/10.48550/arXiv.1512.02325

LMSYS. (2025). Leaderboard Overview. https://lmarena.ai/leaderboard

Lo, K., Wang, L. L., Neumann, M., Kinney, R., & Weld, D. S. (2020). S2ORC: The Semantic Scholar Open Research Corpus. ACL. https://aclanthology.org/2020.acl-main.447/

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction. In ECDL 2009 (pp. 473-475). ResearchGate. https://www.researchgate.net/publication/221176095_GROBID_Combining_Automatic_Bibliographic_Data_Recognition_and_Term_Extraction_for_Scholarship_Publications

Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2021). Web Scraping Techniques and Applications: A Literature Review. SCRS CONFERENCE PROCEEDINGS on INTELLIGENT SYSTEMS, 381-394. https://doi.org/10.52458/978-93-91842-08-6-38

Lou, C., & Yuan, S. (2019). Influencer marketing: How message value and credibility affect consumer trust of branded content on social media. Journal of interactive advertising, 19(1), 58-73. https://doi.org/10.1080/15252019.2018.1533501

Lou, C., Tan, S. S., & Chen, X. (2019). Investigating consumer engagement with influencer-vs. brand-promoted ads: The roles of source and disclosure. Journal of Interactive Advertising, 19(3), 169-186. https://doi.org/10.1080/15252019.2019.1667928

Lyndyuk, A., Havrylyuk, I., Khirivskyi, R., & Kohut, M. (2024). The impact of artificial intelligence on marketing communications: New business opportunities and challenges. Economics of Development 23(4), 60-71. http://dx.doi.org/10.57111/econ/4.2024.60

Mahoney, J., Le Louvier, K., Lawson, S., Bertel, D., & Ambrosetti, E. (2022). Ethical considerations in social media analytics in the context of migration: lessons learned from a Horizon 2020 project. Research Ethics, 18(3). https://doi.org/10.1177/17470161221087542

Malyavkina, I. (2018). Linguistic peculiarities of media texts of financial sphere in the social media space. Traektoriâ nauki, 4(12), 4001-4005. https://www.ceeol.com/search/article-detail?id=728135

Mammassis, C. (2025). The role of digital marketing in building brand awareness in the modern era. British Journal of Management and Marketing Studies, 8(2), 109-125. https://doi.org/10.52589/bjmms_xg92rnys

Marchese Robinson, R. L., Palczewska, A., Palczewski, J., & Kidley, N. (2017). Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. Journal of chemical information and modeling, 57(8), 1773-1792.

Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, *23*(1), 193-201.

Martínez-López, F. J., Anaya-Sánchez, R., Fernández Giordano, M., & Lopez-Lopez, D. (2020). Behind influencer marketing: key marketing decisions and their effects on followers' responses. Journal of Marketing Management, 36(7-8), 579-607. https://doi.org/10.1080/0267257X.2020.1738525

Martinez-Ríos, E. A., Bustamante-Bello, M. R., & Arce-Sáenz, L. A. (2022). A review of road surface anomaly detection and classification systems based on vibration-based techniques. *Applied Sciences*, *12*(19), 9413.

Mason, R., & Clarke, J. (2025). Social media as a compliance risk for financial services: Exploring emerging risks and finding solutions to mitigate harm. Journal of Financial Compliance, 8(3), 245-256. https://doi.org/10.69554/TDHS1278

Masuda, H., Han, S. H., & Lee, J. (2022). Impacts of influencer attributes on purchase intentions in social media influencer marketing: Mediating roles of characterizations. Technological Forecasting and Social Change, 174, 121246. https://doi.org/10.1016/j.techfore.2021.121246

McLeod, S. (2025, March 18). Albert Bandura's social learning theory. Simply Psychology. https://www.simplypsychology.org/bandura.html

Meer, D., & Staubach, K. (2020). Social media influencers' advertising targeted at teenagers: The multimodal constitution of credibility. Visualizing Digital Discourse: International, institutional und ideological perspectives. Berlin, Boston: de Gruyter, 245-269.

Mekhalfi, M. L., Nicolò, C., Bazi, Y., Al Rahhal, M. M., Alsharif, N. A., & Al Maghayreh, E. (2021). Contrasting YOLOv5, transformer, and EfficientDet detectors for crop circle detection in the desert. IEEE Geoscience and Remote Sensing Letters, 19, 1-5. http://dx.doi.org/10.1109/LGRS.2021.3085139

Merkley, K. J., Pacelli, J., Piorkowski, M., & Williams, B. (2024). Crypto-influencers. Review of Accounting Studies, 29(3), 2254-2297. https://doi.org/10.1007/s11142-024-09838-4

Meusel, R., Vigna, S., Lehmberg, O., & Bizer, C. (2014). Graph structure in the web --- revisited. Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, 427-432. https://doi.org/10.1145/2567948.2576928

Mi, A., Huang, X., Huo, Z., & Liu, L. (2025). Context-aware learning and background activation suppression for weakly supervised semantic segmentation. Multimedia Systems, 31(2), 1-12. http://dx.doi.org/10.21203/rs.3.rs-4907075/v1

Michel, A. E., Miller, E. S., Singh, P., Schulz, G., & Limaye, R. J. (2024). The emerging landscape of social media influencers in public health collaborations: A scoping review. Health Promotion Practice, 25(3), 245-262. https://doi.org/10.1177/15248399241258442

Miettinen, M. (2025). The Impact of Social Media on Investors' Risk-Taking. https://urn.fi/URN:NBN:fi-fe2025052755186

Mikelionytė, M., & Lezgovko, A. (2021). Gender impact on personal investment strategies. Economics and Culture, 18(1), 32-45. https://doi.org/10.2478/jec-2021-0003

Milligan, I. (2017). Welcome to the web: The online community of GeoCities during the early years of the World Wide Web. The Web as History. Using Web Archives to Understand the Past and the Present, 137-158. https://doi.org/10.25969/mediarep/12521

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Mitchell, R. (2018). Web Scraping with Python. "O'Reilly Media, Inc."

Mittal, D. (2024). Regulating likes and listings: The rise of the finfluencer and the implications for securities law. American Bar Association. https://cba.org/sections/business-law/resources/regulating-likes-and-listings-the-rise-of-the-finfluencer-and-the-implications-for-securities-law-i/

Mittal, P. (2024). A comprehensive survey of deep learning-based lightweight object detection models for edge devices. Artificial Intelligence Review, 57(9), 242-257. https://doi.org/10.1007/s10462-024-10877-1

Mohammed, S. Y. (2025). Architecture review: Two-stage and one-stage object detection. Franklin Open, 100-122. https://doi.org/10.1016/j.fraope.2025.100322

Mondschein, C. F., & Monda, C. (2018). The EU's General Data Protection Regulation (GDPR) in a Research Context. Fundamentals of Clinical Data Science, 55-71. https://doi.org/10.1007/978-3-319-99713-1_5

Mongeon, P., & Paul-Hus, A. (2016). The Journal Coverage of Web of Science and Scopus: a Comparative Analysis. Scientometrics, 106(1), 213-228. https://doi.org/10.1007/s11192-015-1765-5

Montag, C., & Hegelich, S. (2020). Understanding detrimental aspects of social media use: will the real culprits please stand up?. Frontiers in Sociology, 5, 599270. https://doi.org/10.3389/fsoc.2020.599270

Moreish. (2025, August 11). The rise of the 'Finfluencer': What FS Marketers need to know. Moreish Marketing. https://moreishmarketing.com/views/the-rise-of-the-finfluencer/

Mölders, M., Bock, L., Barrantes, E., & Zülch, H. (2024). Exploring the emergence of finfluencers on Instagram: Survey-based descriptive insights into their segmentation, motivations, business models, and potential use in financial communication (pp. 1-42). HHL Leipzig Graduate School of Management. https://doi.org/10.2139/ssrn.4971402

Mölders, M., Bock, L., Barrantes, E., & Zülch, H. (2025). Understanding finfluencers: Roles and strategic partnerships in retail investor engagement. Journal of Business Research, 198, 115462. https://doi.org/10.1016/j.jbusres.2025.115462

Murat, A. A., & Kiran, M. S. (2025). A comprehensive review of YOLO versions for object detection. Engineering Science and Technology, an International Journal, 70, 102-161. https://doi.org/10.1016/j.jestch.2025.102161

Mutlu, M. A., Ulku, E. E., & Yildiz, K. (2024). A web scraping app for smart literature search of the keywords. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/peerj-cs.2384

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A Step-by-Step process of thematic analysis to develop a conceptual model in qualitative research. International Journal of Qualitative Methods, 22. https://doi.org/10.1177/16094069231205789

Naeem, M., Smith, T., & Thomas, L. (2025). Thematic Analysis and Artificial intelligence: A Step-by-Step process for using ChatGPT in thematic analysis. International Journal of Qualitative Methods, 24. https://doi.org/10.1177/16094069251333886

Najork, M., & Wiener, J. L. (2001). Breadth-first crawling yields high-quality pages. https://doi.org/10.1145/371920.371965

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. 2008 IEEE Symposium on Security and Privacy (SP 2008), 1(1). https://doi.org/10.1109/sp.2008.33

Nasser, M., Arshad, N. I., Ali, A., Alhussian, H., Saeed, F., Da'u, A., & Nafea, I. (2025). A systematic review of multimodal fake news detection on social media using Deep Learning Models. *Results in Engineering*, *26*, 104752. https://doi.org/10.1016/j.rineng.2025.104752

Neha, F., Bhati, D., Shukla, D. K., & Amiruzzaman, M. (2025). From classical techniques to convolution-based models: A review of object detection algorithms. In 2025, IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS) (pp. 1-6). IEEE. https://arxiv.org/html/2412.05252v1

Neha, Y., Saritha, V., Samyuktha, N., Gayathri, B., & Charith, A. (2022). Smart parking system using object detection. In Proceedings of the 2nd Indian International Conference on Industrial Engineering and Operations Management (pp. 148-154).

Nguyen, T., Park, E. A., Han, J., Park, D. C., & Min, S. Y. (2014). Object detection using scale-invariant feature transform. In Genetic and Evolutionary Computing: Proceedings of the Seventh International Conference on Genetic and Evolutionary Computing, ICGEC 2013, August 25-27, 2013, Prague, Czech Republic (pp. 65-72). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-01796-9_7

Nissenbaum, H. F. (2010). Privacy in context: technology, policy, and the integrity of social life. Stanford University Press.

Noyes, J., Booth, A., Moore, G., Flemming, K., Tunçalp, Ö., & Shakibazadeh, E. (2019). Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs, and outlining some methods. BMJ Global Health, 4(Suppl 1), e000893. https://doi.org/10.1136/bmjgh-2018-000893

OECD (2021). *Financial Literacy Levels in the Commonwealth of Independent States in 2021*. OECD Publishing. https://doi.org/10.1787/394dbe88-en.

Ogunleye, B., Sharma, H., & Shobayo, O. (2024). Sentiment-informed sentence Bert-ensemble algorithm for Depression Detection. *Big Data and Cognitive Computing*, *8*(9), 112. https://doi.org/10.3390/bdcc8090112

Okorie, N. G. N., Udeh, N. C. A., Adaga, N. E. M., DaraOjimba, N. O. D., & Oriekhoe, N. O. I. (2024). Ethical considerations in data collection and analysis: A review: Investigating ethical practices and challenges in modern data collection and analysis. International Journal of Applied Research in Social Sciences, 6(1), 1-22. https://doi.org/10.51594/ijarss.v6i1.688

Ollama. (2024). OpenAI compatibility. https://ollama.com/blog/openai-compatibility

Ollama. (2025). Ollama's documentation. https://docs.ollama.com/

Olston, C., & Najork, M. (2010). Web Crawling. Foundations and Trends® in Information Retrieval, 4(3), 175-246. https://doi.org/10.1561/1500000017

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

Padilla, R., Passos, W. L., Dias, T. L., Netto, S. L., & Da Silva, E. A. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, *10*(3), 279.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

Park, K., Zhou, T., & D'Antoni, L. (2025). Flexible and efficient grammar-constrained decoding. *arXiv preprint arXiv:2502.05111*.

Park, L. L. (2025). Hiq Labs v. LinkedIn Case Study and Its Implication for the Open Banking Regulation in the US Compared with the Current Cases and Regulations in South Korea. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5350952

Parry, M. (2011). Harvard Researchers Accused of Breaching Students' Privacy. Chronicle.com. https://www.chronicle.com/article/harvard-researchers-accused-of-breaching-students-privacy/

Patnaik, S. K., & Narendra Babu, C. (2021). Trends in web data extraction using machine learning. *Web Intelligence*, *19*(3), 169–190. https://doi.org/10.3233/web-210465

Pebrianto, W., Mudjirahardjo, P., Pramono, S. H., & Setyawan, R. A. (2023). YOLOv3 with spatial pyramid pooling for object detection with unmanned aerial vehicles. arXiv preprint

arXiv:2305.12344, 1-6 https://www.researchgate.net/publication/370949055_YOLOv3_with_Spatial_Pyramid_Pooling_for_Object_Detection_with_Unmanned_Aerial_Vehicles

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition, 188, 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590-595. https://doi.org/10.1038/s41586-021-03344-2

Pinto-Coelho, L. (2023). How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering, 10(12), 14-35. https://doi.org/10.3390/bioengineering10121435

Pittman, M., & Abell, A. (2021). More trust in fewer followers: Diverging effects of popularity metrics and green orientation social media influencers. Journal of Interactive Marketing, 56(1), 70-82. https://doi.org/10.1016/j.intmar.2021.05.002

Plachouras, V., Carpentier, F., Faheem, M., Masanès, J., Risse, T., Senellart, P., Siehndel, P., & Stavrakas, Y. (2014). ARCOMEM Crawling Architecture. Future Internet, 6(3), 518-541. https://doi.org/10.3390/fi6030518

Pokhrel, L., Bhattarai, P., & Krishna Pokhrel, S. (2025). Are Financial Influencers Helping us with Financial Decision-Making? An Application of Structural Equation Modeling and Artificial Neural Networking Approach. Journal of Promotion Management, 31(3), 485-514. https://doi.org/10.1080/10496491.2025.2466591

Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-45914-8

Politz, D. (2024, December 13). Reflexive thematic analysis (RTA) in qualitative research. Delve. https://delvetool.com/blog/reflexive-thematic-analysis

Putra, H. A. A., Murni, A., & Chahyati, D. (2025). Enhancing Bounding Box Regression for Object Detection: Dimensional Angle Precision IoU-Loss. IEEE Access. http://dx.doi.org/10.1109/ACCESS.2025.3567767

Qwen Team. (2024). Qwen2.5 technical report. arXiv. https://arxiv.org/abs/2407.10671

Raamkumar, A. S., Tan, S. G., & Wee, H. L. (2020). Measuring the Outreach Efforts of Public Health Authorities and the Public Response on Facebook During the COVID-19 Pandemic in Early 2020: Cross-Country Comparison (Preprint). Journal of Medical Internet Research. https://doi.org/10.2196/19334

Rachmad, Y. E. (2024). The Future of Influencer Marketing: Evolution of Consumer Behavior in the Digital World. PT. Sonpedia Publishing Indonesia.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Computer Vision and Pattern Recognition*.

Rajkumar, C. (2025). Influencer Marketing on Brand Awareness and Purchase Decision Among Gen z and Millennials. Journal of Marketing & Social Research, 2, 164-170. https://doi.org/10.61336/jmsr/25-07-20

Reason, T., Langham, J., & Gimblett, A. (2024). Automated mass extraction of over 680,000 PICOs from clinical study abstracts using generative AI: a proof-of-concept study. *Pharmaceutical Medicine*, *38*(5), 365-372.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, 779-788. https://doi.org/10.1109/CVPR.2016.91

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. ResearchGate. https://www.researchgate.net/publication/336996965_Sentence-BERT_Sentence_Embeddings_using_Siamese_BERT-Networks

Reinikainen, H., Munnukka, J., Maity, D., & Luoma-Aho, V. (2020). 'You really are a great big sister' parasocial relationships, credibility, and the moderating role of audience comments in influencer marketing. Journal of marketing management, 36(3-4), 279-298. https://doi.org/10.1080/0267257X.2019.1708781

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 1-14. https://doi.org/10.48550/arXiv.1506.01497

Rietveld, R., Van Dolen, W., Mazloom, M., & Worring, M. (2020). What you feel, is what you like influence of message appeals on customer engagement on Instagram. Journal of interactive marketing, 49(1), 20-53. https://doi.org/10.1016/j.intmar.2019.06.003

Rogers, R. (2013). Digital methods. The MIT Press.

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. Palgrave Communications, 5(1), 1-10. https://doi.org/10.1057/s41599-019-0279-9

Rosero, L. A., Gomes, I. P., da Silva, J. A., Przewodowski, C. A., Wolf, D. F., & Osório, F. S. (2024). Integrating modular pipelines with end-to-end learning: A hybrid approach for robust and reliable autonomous driving systems. *Sensors*, *24*(7), 2097. https://doi.org/10.3390/s24072097

Roziewski, S., & Kozłowski, M. (2021). LanguageCrawl: a generic tool for building language models upon the common crawl. Language Resources and Evaluation, 55(4), 1047-1075. https://doi.org/10.1007/s10579-021-09551-7

Sabrin, A., Alviansyah, A., & Anca, A. (2025). The ethics of influencer marketing: Transparency, trust, and consumer perceptions. Social Science Research Network, 1-28. https://doi.org/10.2139/ssrn.5112872

Sahni, N. S. (2016). Advertising spillovers: Evidence from online field experiments and implications for returns on advertising. Journal of Marketing Research, 53(4), 459-478. https://doi.org/10.1509/jmr.14.0274

Salaris, S., Ocagli, H., Casamento, A., Lanera, C., & Gregori, D. (2025). Foodborne event detection based on Social Media Mining: A Systematic Review. *Foods*, *14*(2), 239. https://doi.org/10.3390/foods14020239

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., & Vespignani, A. (2012). Digital Epidemiology. PLoS Computational Biology, 8(7), e1002616. https://doi.org/10.1371/journal.pcbi.1002616

Samuelson, P. (2013). Aaron Swartz: Opening Access to Knowledge. UC Berkeley Law. https://www.law.berkeley.edu/article/aaron-swartz-opening-access-to-knowledge/

Sapkota, R., Qureshi, R., Calero, M. F., Badjugar, C., Nepal, U., Poulose, A., & Karkee, M. (2024). Yolov12 to its genesis: A decadal and comprehensive review of the You Only Look Once (yolo) series. arXiv preprint arXiv:2406.19407. https://doi.org/10.1007/s10462-025-11253-3

Schivinski, B., & Dabrowski, D. (2016). The effect of social media communication on consumer perceptions of brands. Journal of Marketing Communications, 22(2), 189-214. https://doi.org/10.1080/13527266.2013.871323

Schmidt, L., Hair, K., Graziosi, S., Campbell, F., Kapp, C., Khanteymoori, A., ... & Thomas, J. (2024). Exploring the use of a large language model for data extraction in systematic reviews: a rapid feasibility study. *arXiv preprint arXiv:2405.14445*.

Schmidt, L., Olorisade, B. K., McGuinness, L. A., Thomas, J., & Higgins, J. P. (2020). Data extraction methods for systematic review (semi) automation: A living review protocol. *F1000Research*, *9*, 210.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29. https://doi.org/10.1177/1536867X20909688

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 1715-1725). Association for Computational Linguistics. https://aclanthology.org/P16-1162/

SEC. U.S. Securities and Exchange Commission. (2022). Title of the 2022 Agency Financial Report. SEC. https://www.sec.gov/files/sec-2022-agency-financial-report.pdf

Shahriar, T. (2025). Comparative Analysis of Lightweight Deep Learning Models for Memory-Constrained Devices. arXiv preprint arXiv:2505.03303. https://arxiv.org/html/2505.03303v1

Shaikh, I. M., Akhtar, M. N., Aabid, A., & Ahmed, O. S. (2024). Enhancing sustainability in the production of palm oil: Creative monitoring methods using YOLOv7 and YOLOv8 for effective plantation management. Biotechnology Reports, 44, 1-13. https://doi.org/10.1016/j.btre.2024.e00853

Shan, Y., Chen, K. J., & Lin, J. S. (2020). When social media influencers endorse brands: The effects of self-influencer congruence, parasocial identification, and perceived endorser motive. International journal of advertising, 39(5), 590-610. https://doi.org/10.1080/02650487.2019.1678322

Sharp, K., Ouellette, R. R., Singh, R. S., DeVito, E. E., Kamdar, N., de la Noval, A., Murthy, D., & Kong, G. (2025). Generative Artificial Intelligence and machine learning methods to screen social media content. *PeerJ Computer Science*, *11*. https://doi.org/10.7717/peerj-cs.2710

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shen, H., Ju, Y., & Zhu, Z. (2023). Extracting useful emergency information from social media: A method integrating machine learning and rule-based classification. *International Journal of Environmental Research and Public Health*, *20*(3), 1862. https://doi.org/10.3390/ijerph20031862

Shetty, N. P., Bijalwan, Y., Chaudhari, P., Shetty, J., & Muniyal, B. (2024). Disaster Assessment from Social Media using multimodal deep learning. *Multimedia Tools and Applications*, *84*(18), 18829–18854. https://doi.org/10.1007/s11042-024-19818-0

Shiller, R. J. (2019). Narrative economics: How stories go viral and drive major economic events. The Quarterly Journal of Austrian Economics, 22(4), 620-627. https://doi.org/10.1080/09672567.2021.1928927

Shin, D. (2022). The actualization of meta-affordances: Conceptualizing affordance actualization in the metaverse games. Computers in human behavior, 133, 107292. https://doi.org/10.1016/j.chb.2022.107292

Singh, A. A. (2025). The Rise of Social Trading: Economics of Influence in Financial Markets. Available at SSRN 5371091. http://dx.doi.org/10.2139/ssrn.5371091

Singh, N., Albishri, N., Galgotia, A., Cillo, V., & Papa, A. (2025). Decoding influence: understanding the dynamics of consumers' engagement with financial influencers on social media. Journal of Enterprise Information Management, 1-23. https://doi.org/10.1108/JEIM-03-2025-0186

Singh, S., & Sarva, M. (2024). The rise of finfluencers: A digital transformation in investment advice. Australasian Accounting, Business and Finance Journal, 18(3). https://doi.org/10.14453/aabfj.v18i3.14

Sokolova, K., & Kefi, H. (2020). Instagram and YouTube bloggers promote it, why should I buy? How credibility and parasocial interaction influence purchase intentions. Journal of retailing and consumer services, 53, 101742. https://doi.org/10.1016/j.jretconser.2019.01.011

Song, G., Liu, Y., & Wang, X. (2020). Revisiting the sibling head in the object detector. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11563-11572. https://doi.org/10.48550/arXiv.2003.07540

Soto-Vásquez, A. D., & Jimenez, N. (2022). A dramaturgical analysis of Latina influencers use of props and settings to signal identity. *Journalism and Media*, *3*(3), 407-418.

Sousa, R., Nogueira, M., & Gomes, S. (2025). How influential can a sustainable digital influencer be? An exploratory study on the relationship between influencing skills and followers' environmental concerns, awareness and pro-environmental behaviour. *International Journal of Innovation Science*.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics - Topics discovery, data collection, and data preparation challenges. International Journal of Information Management, 39(39), 156-168. https://www.sciencedirect.com/science/article/pii/S0268401217308526

Stubb, C., Nyström, A. G., & Colliander, J. (2019). Influencer marketing: The impact of disclosing sponsorship compensation justification on sponsored content effectiveness. Journal of Communication Management, 23(2), 109-122. https://doi.org/10.1108/JCOM-11-2018-0119

Subagio, H., Satoto, S., & Ediningsih, S. (2021). The effect of investment education and investment experience on investment decision with financial knowledge as an intervening variable. Russian Journal of Agricultural and Socio-Economic Sciences, 99(3), 143-150. https://doi.org/10.18551/rjoas.2020-03.16

Sumit, S. B., Joshi, S., & Rana, U. (2024). Comprehensive review of R-CNN and its variant architectures. Int. Res. J. Adv. Eng. Hub IRJAEH, 2(04), 959-966. http://dx.doi.org/10.47392/IRJAEH.2024.0134

Sun, Y., Sun, Z., & Chen, W. (2024). The evolution of object detection methods. Engineering Applications of Artificial Intelligence, 133, 1-17. https://doi.org/10.1016/j.engappai.2024.108458

Sun, Z., Zhang, R., Doi, S. A., Furuya-Kanamori, L., Yu, T., Lin, L., & Xu, C. (2024). How good are large language models for automated data extraction from randomized trials?. *MedRxiv*, 2024-02.

Sundarasen, S., Rajagopalan, U., Kanapathy, M., & Kamaludin, K. (2023). Women's financial literacy: A bibliometric study on current research and future directions. Heliyon, 9(12), e21379. https://doi.org/10.1016/j.heliyon.2023.e21379

Surugiu, M., Vasile, V., Surugiu, C., Mazilescu, C. R., Panait, M., & Bunduchi, E. (2025). Tax Compliance Pattern Analysis: A Survey-Based Approach. International Journal of Financial Studies, 13(1), 14. https://doi.org/10.3390/ijfs13010014

Symbiosis, A. R., & Gandhi, A. (2024, September). Finfluencer: Exploring the untapped influence of financial influencers. In 2024 14th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 190-196). IEEE. DOI: 10.1109/ACIT62333.2024.10712618

Tafesse, W., & Wood, B. P. (2021). Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy. Journal of retailing and consumer services, 58, 102303. https://doi.org/10.1016/j.jretconser.2020.102303

Tan, L., Huangfu, T., Wu, L., & Chen, W. (2021). Comparison of yolo v3, Faster R-CNN, and SSD for real-time pill identification. 1-9. http://dx.doi.org/10.21203/rs.3.rs-668895/v1

Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.01079

Tao, D., Hu, R., Zhang, D., Laber, J., Lapsley, A., Kwan, T., Rathke, L., Rundensteiner, E., & Feng, H. (2023). A novel foodborne illness detection and web application tool based on social media. *Foods*, *12*(14), 2769. https://doi.org/10.3390/foods12142769

Taunk, P., Jayasri, G., Priya, J. P., & Kumar, N. S. (2020). Face detection using Viola-Jones with Haar cascade. Test Engineering and Management, 83(06). 19146-19149. https://www.researchgate.net/publication/342122717_Face_Detection_using_Viola_Jones_with_Haar_Cascade

Terry, G., & Hayfield, N. (2021). Reflexive thematic analysis. In Handbook of research methods in social and political science (pp. 430-441). Edward Elgar Publishing. https://doi.org/10.4337/9781788977159.00049

Evans, N. J., Phua, J., Lim, J., & Jun, H. (2017). Disclosing Instagram influencer advertising: The effects of disclosure language on advertising recognition, attitudes, and behavioral intent. *Journal of interactive advertising*, *17*(2), 138-149.

Thelwall, M. (2001). Extracting macroscopic information from Web links. Journal of the American Society for Information Science and Technology, 52(13), 1157-1168. https://doi.org/10.1002/asi.1182

Thelwall, M. (2008). Bibliometrics to webometrics. Journal of Information Science, 34(4), 605-621. https://doi.org/10.1177/0165551507087238

Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2021). Algorithmic inference, political interest, and exposure to news and politics on Facebook. Information, Communication & Society, 24(2), 183-200. https://doi.org/10.1080/1369118X.2019.1642934

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, Ł. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, *18*(4), 317-335.

Tomșa, M., Romonți-Maniu, A., & Scridon, M. (2021). Is Sustainable Consumption Translated into Ethical Consumer Behavior? Sustainability, 13(6), 3466. https://doi.org/10.3390/su13063466

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trinugroho, I., & Sembel, R. (2011). Overconfidence and Excessive trading Behavior: an experimental study. International Journal of Business and Management, 6(7). https://doi.org/10.5539/ijbm.v6n7p147

Tsvetkova, M., Garcia-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even Good Bots Fight: The Case of Wikipedia. SSRN Electronic Journal. http://dx.doi.org/10.48550/arXiv.1609.04285

Tufekci, Z. (2014). SOCIAL MOVEMENTS AND GOVERNMENTS IN THE DIGITAL AGE: EVALUATING A COMPLEX LANDSCAPE. Journal of International Affairs, 68(1), 1-18. http://www.jstor.org/stable/24461703

Ultralytics. (2025). YOLO12: Rilevamento di oggetti incentrato sull'attenzione. Retrieved from https://docs.ultralytics.com/it/models/yolo12/

Urban, J. M., & Quilter, L. (2016). Efficient Process or "Chilling Effects"? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act. Bepress Legal Repository (Bepress (United States)), 22(4). https://doi.org/10.31235/osf.io/pyzua

Van Dam, S., & Van Reijmersdal, E. (2019). Insights in adolescents' advertising literacy, perceptions and responses regarding sponsored influencer videos and disclosures. Cyberpsychology: journal of psychosocial research on cyberspace, 13(2).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vikan, M., Aryan, R., Kannelønning, M. S., Riegler, M. A., & Danielsen, S. O. (2025). Reflecting on LLM support in reflexive thematic analysis: An exploratory study. Qualitative Health Research, 35(2), 112-128. https://doi.org/10.1177/10497323251365211

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science (New York, N.Y.), 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Vrontis, D., Makrides, A., Christofi, M., & Thrassou, A. (2021). Social media influencer marketing: A systematic review, integrative framework and future research agenda. International Journal of Consumer Studies, 45(4), 617-644. https://doi.org/10.1111/ijcs.12647

Wahyudi, M. A., Rahmadhani, M. V., Mu'is, A., & Evelyna, F. (2025). The impact of Short-Form video marketing, influencer relatability, and trust signals on Gen Z's purchase intention. International Journal of Business Law and Education, 6(1), 855-864. https://doi.org/10.56442/ijble.v6i1.1108

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *2023 IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 7464–7475. https://doi.org/10.1109/cvpr52729.2023.00721

Wang, D., & Wu, H. (2021). IoU regression with H+ L-sampling for accurate detection confidence. Sensors, 21(13), 4433-4437. https://doi.org/10.3390/s21134433

Wang, T., & Li, Y. (2022). Enhanced Task-Aware Spatial Disentanglement Head for Oil Tanks Detection in High-Resolution Optical Imagery. IEEE Geoscience and Remote Sensing Letters, 19, 1-5. http://dx.doi.org/10.1109/LGRS.2022.3184836

Warkulat, S., & Pelster, M. (2024). Social media attention and retail investor behavior: Evidence from r/wallstreetbets. International Review of Financial Analysis, 96, 103721. https://doi.org/10.1016/j.irfa.2024.103721

Wei, J., As'arry, A., Rezali, K. A. M., Yusoff, M. Z. M., Ma, H., & Zhang, K. (2025). A Review of the YOLO Algorithm and Its Applications in Autonomous Driving Object Detection. IEEE Access. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11015428

Wei, X., Kumar, N., & Zhang, H. (2025). Addressing Bias in Generative AI: Challenges and research opportunities in information Management. Information & Management, 62(2), 104103. https://doi.org/10.1016/j.im.2025.104103

Weigle, M. C. (2023). The Use of Web Archives in Disinformation Research. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2306.10004

Wellman, M. L., Stoldt, R., Tully, M., & Ekdale, B. (2020). Ethics of authenticity: Social media influencers and the production of sponsored content. Journal of media ethics, 35(2), 68-82. https://doi.org/10.1080/23736992.2020.1736078

Wittmann, F. H. (2024). Enhancing thematic analysis with large language models: A comparative study of structured prompting techniques (pp. 1-85). DiVA Portal. https://www.diva-portal.org/smash/get/diva2:1939104/FULLTEXT02

Wojdynski, B. W. (2016). Native advertising: Engagement, deception, and implications for theory. The new advertising: Branding, content and consumer relationships in a data-driven social media era, 203-236.

Wong, L. Z., Bhattacharya, P., Loh, S. B., Oh, H. S., Juraimi, S. A., Anant, N., Pillay, A., Chong, M. F.-F., Fogel, A., Sheen, F., & Pink, A. E. (2025, March 5). Utilizing large language models to conduct thematic analysis: A case study on focus group transcripts. SSRN. https://doi.org/10.2139/ssrn.5167505

Woody, S. K., Burdick, D., Lapp, H., & Huang, E. S. (2020). Application programming interfaces for knowledge transfer and generation in the life sciences and healthcare. Npj Digital Medicine, 3(1). https://doi.org/10.1038/s41746-020-0235-5

World Economic Forum. (2025). Are “finfluencers” the future of financial advice worldwide? https://www.weforum.org/stories/2024/07/finfluencer-financial-advice-social-media/

World Health Organization. (2021). Building a response workforce to manage infodemics: WHO competency framework (Publication No. 9789240035287). World Health Organization. https://www.who.int/publications/i/item/9789240035287

Wu, S., Li, X., & Wang, X. (2020). IoU-aware single-stage object detector for accurate localization. Image and Vision Computing, 97, 1-9. http://dx.doi.org/10.1016/j.imavis.2020.103911

Xin, Z., Chen, S., Wu, T., Shao, Y., Ding, W., & You, X. (2024). Few-shot object detection: Research advances and challenges. Information Fusion, 107, 102307. https://arxiv.org/html/2404.04799v1

Xu, X., & Pratt, S. (2018). Social media influencers as endorsers to promote travel destinations: an application of self-congruence theory to the Chinese Generation Y. Journal of travel & tourism marketing, 35(7), 958-972. https://doi.org/10.1080/10548408.2018.1468851

Yen, C. S., Chen, G. L., Kang, C. C., Wang, Y. H., & Yang, S. C. (2024). Investigating factors that influence purchase intentions in live-streaming contexts through the elaboration likelihood model: the perspectives of para-social interaction and information quality. The Review of Socionetwork Strategies, 18(2), 387-406. https://doi.org/10.1007/s12626-024-00166-2

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, *11*(12). https://doi.org/10.1093/nsr/nwae403

Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck. org and FlackCheck. org. Journalism & Mass Communication Quarterly, 95(1), 49-75. https://doi.org/10.1177/1077699017710453

Yu, J., Bekerian, D. A., & Osback, C. (2024). Navigating the digital landscape: challenges and barriers to effective information use on the internet. Encyclopedia, 4(4), 1665-1680. https://doi.org/10.3390/encyclopedia4040109

Yu, L., Tang, L., & Mu, L. (2025). A Review of Detection Transformer: From Basic Architecture to Advanced Developments and Visual Perception Applications. Sensors (Basel, Switzerland), 25(13), 1-51. https://doi.org/10.3390/s25133952

Yılmazdoğan, O. C., Doğan, R. Ş., & Altıntaş, E. (2021). The impact of the source credibility of Instagram influencers on travel intention: The mediating role of parasocial interaction. Journal of Vacation Marketing, 27(3), 299-313. https://doi.org/10.1177/1356766721995973

Zachlod, C., Samuel, O., Ochsner, A., & Werthmüller, S. (2022). Analytics of social media data - State of characteristics and application. Journal of Business Research, 144, 1064-1076. https://doi.org/10.1016/j.jbusres.2022.02.016

Zhang, H., & Gong, X. (2021). Consumer susceptibility to social influence in new product diffusion networks: how does network location matter? European Journal of Marketing, 55(5), 1469-1488. https://doi.org/10.1108/EJM-06-2019-0491

Zhang, Z., Ma, W., & Vosoughi, S. (2024). Is GPT-4V (ISION) all you need for automating academic data visualization? exploring vision-language models' capability in reproducing academic charts. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8271–8288. https://doi.org/10.18653/v1/2024.findings-emnlp.485

Zhang, X., Razavi-Far, R., Isah, H., David, A., Higgins, G., Lu, R., & Ghorbani, A. A. (2024). Area in circle: A novel evaluation metric for object detection. Knowledge-Based Systems, 293, 111-168. https://doi.org/10.1016/j.knosys.2024.111684

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., & Ling, H. (2019). M2det: A single-shot object detector based on a multi-level feature pyramid network. In Proceedings of the

AAAI conference on artificial intelligence, 33(1), 9259-9266. https://doi.org/10.1609/aaai.v33i01.33019259

Zhou, P., Ni, B., Geng, C., Hu, J., & Xu, Y. (2018). Scale-transferrable object detection. In proceedings of the IEEE conference on computer vision and pattern recognition, 528-537. https://doi.org/10.1109/cvpr.2018.00062

Zhu, R., Zhang, S., Wang, X., Wen, L., Shi, H., Bo, L., & Mei, T. (2019). ScratchDet: Training single-shot object detectors from scratch. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2268-2277. https://doi.org/10.48550/arXiv.1810.08425

Zimmer, M. (2010). "But the data is already public": on the research ethics in Facebook. Ethics and Information Technology, 12(4), 313-325. https://doi.org/10.1007/s10676-010-9227-5

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.

Appendix A: Thematic Analysis Summary

Table A1 *Summary of Thematic Analysis Output*

Theme ID/Name	Keywords	Codes
T1 Women-First Money Empowerment (Financial Feminism)	"Financial Feminists", "builder girlie", "anti-shame", "boundaries", "safety", "autonomy", "equity", "solidarity", "community", "we/our"	• Feminist Pedagogy • Safe Money Space • Boundary Setting • Equity Mindset
T2 Budgeting & Cashflow Systems	"Budget-by-Paycheck (BBP)", "cash envelopes", "zero-based", "sinking funds", "track → interpret", "automation", "Saveopoly"	• Paycheck Mapping • Envelope Buckets • Leak Pruning • Auto-Pilot
T3 Emergency Funds & HYSA (Liquidity)	"starter EF", "3-6 months", "HYSA", "cash management", "rebuild after use", "HYSA100"	• Starter Cushion • HYSA Parking • Rebuild Ritual
T4 Debt Reduction Systems	"Avalanche", "Snowball", "APR order", "consolidation", "debt-free milestones", "relapse avoidance"	• Avalanche Sequencing • Snowball Momentum • Disciplined Consolidation • Relapse Guardrails
T5 Credit Health & Cost of Capital	"utilization", "on-time pay", "age of accounts", "disputes", "secured cards", "AU tradelines", "report vs score"	• Score Hygiene • Thin-File On-Ramps • Report Mastery

		• Fee Avoidance
T6 Credit Cards, Rewards & Travel Hacking	"category multipliers", "minimum spend", "redemptions", "lounges", "CARDRECS", "avoid store-card traps"	• Rewards Stack • Redemption Literacy • Guardrails First
T7 Retirement & Account Hygiene	"employer match", "Roth IRA", "HSA", "catch-ups", "rollovers", "orphan 401(k)s", "RTI/Capitalize", "April 15"	• Contribution Order • Rollover Cleanup • Deadline Discipline
T8 Investing 101 & Long-Termism	"index funds", "ETFs", "diversification", "DCA", "buy the whole market", "time-in-market"	• Index Default • DCA Automation • Long-Horizon Mindset
T9 Taxes & Paycheck Optimization	"W-4 tuning", "stop loaning IRS", "free tax prep", "credits", "SE tax", "quarterly estimates"	• Withholding Hygiene • Access to Filing • SE Compliance
T10 Insurance & Protection + Estate Basics	"term > whole", "disability", "umbrella", "beneficiaries", "will/trust", "estate bundles"	• Right-Sized Transfer • Income Protection • Beneficiary/Estate Hygiene
T11 Housing & Real Estate Literacy	"rent vs buy", "pre-approval", "down payment", "PMI", "closing costs", "maintenance", "don't rush the house"	• Ownership Readiness • Lifetime Costing • Entry Tactics • Debt-Tool Discipline

T12 Income Growth: Career Capital	“negotiation scripts”, “banned questions”, “resume bullets”, “portfolios”, “benefits as comp”, “PTO vs salary”	• Negotiation Scripts • Artifact Quality • Benefit Levers
T13 Side Hustles & Entrepreneurship	“UGC deals”, “first $1,500”, “LLC/S-Corp”, “bookkeeping”, “invoicing”, “pricing”, “repeat-need products”, “comment READY”	• Client On-Ramps • Entity/Admin Hygiene • Pricing Power
T14 Creator Economy & Audience Growth	“hooks”, “social proof”, “email list”, “deliverability”, “platform risk”, “DM codes”, “free intro classes”, “workshop tours”	• Growth Mechanics • Owned Audience • Platform Hedge • Live Education Onramps
T15 CTAs, Lead Magnets & Community Offers	“QUIZ”, “HYSA100”, “CARDRECS”, “TABLE136/169/RTI”, “bootcamps”, “webinars”, “membership/Society”, “DM 'BLUEPRINT'“, “Telegram/Discord rooms”	• Comment-to-Get • Live Events • Cohort/Membership • Community Rooms
T16 Advocacy & Civic Issues via Money	“abortion is a financial issue”, “donation matches”, “voter info”, “Black wealth panels”, “community calls”	• Pocketbook Policy • Access Campaigns • Community Panels

T17 Money & Relationships	"prenups", "provider/protector roles", "wedding costs", "financial infidelity", "shared vs separate", "money dates"	• Prenup Literacy • Roles Discourse • System Design
T18 Parenting, Kids & Intergenerational Planning	"529", "UTMA/UGMA", "teen credit on-ramps", "baby budgets", "childcare", "single-parent credits"	• Kid Accounts • Teen On-Ramps • Family Costing
T19 Values, Mindset & Mental Health	"discipline > motivation", "accountability partners", "money dates", "anxiety", "burnout", "planned rest"	• Habits Over Hacks • Accountability Loops • Anti-Burnout Cadence
T20 Consumer Savviness & Frugality	"subscription audit", "renegotiate", "razor-and-blade model", "upsells", "low-effort swaps", "$3 coffee hacks"	• Bill Pruning • Model Literacy • Everyday Swaps
T21 Safety, Fraud & OPSEC	"identity theft", "freezes", "monitoring", "phishing", "mortgage-fraud tales", "chargebacks", "blacklists", "deepfakes", "impersonation", "platform outages"	• Identity Hygiene • Fraud Case Learning • Transaction Controls • Social-Engineering Defense
T22 Small-Business Finance & Cost Structure	"FTE cost", "layoffs", "contractors", "SE tax",	• People-Cost Reality • SE Tax Hygiene • Spend Scrubbing

	"receipts", "expense audits", "cash runway"	
T23 Women of Color / First-Gen Wealth Narratives	"first-gen Latina", "representation", "rewrite scripts", "approachable basics", "access"	• Representation Matters • Script Rewriting • Access On-Ramps
T24 Health × Money	"screenings", "free annual visits", "GLP-1/Ozempic costs", "relapse economics", "fitness"	• Preventive ROI • Fitness Economics • Drug Cost/Relapse
T25 Housing Access & Affordability	"roommates", "move home briefly", "city cost maps", "homes too expensive", "bridge tactics"	• Bridge Tactics • Temporary Tradeoffs • Market Sentiment
T26 Market Contexts & Economic Narratives	"inflation", "recession", "Fed", "rate hikes", "market volatility", "jobs data"	• Macro Framing • Narrative Calibration • Emotional Anchoring
T27 Tech Tools & Financial Apps	"budget apps", "automation", "AI assistants", "fintech dashboards", "data security"	• Tool Literacy • Automation Flow • Integration Hygiene
T28 Sustainability & Ethical Finance	"ESG", "green investing", "values-based portfolios", "impact investing"	• Impact Alignment • Ethical Screening • Conscious Capital

T29 Behavioral Economics & Nudges	"default settings", "framing effects", "loss aversion", "commitment devices"	• Choice Architecture • Friction Design • Reward Loops
T30 Faith, Purpose & Money Philosophy	"stewardship", "giving", "gratitude", "tithing"	• Stewardship Framing • Gratitude Economics • Purpose Anchors
T31 Trader Education & Execution	"pips", "stop-loss", "risk management", "candlestick entries", "mentor programs", "beginner workshops"	• Core Concepts • Risk Habits • Beginner Onramps
T32 Fintech Infrastructure & Market Access	"event contracts", "prediction markets", "stablecoin/legal rails", "exchange hour extensions", "custody", "support"	• New Products • TradFi-Crypto Bridges • Access & Support
T33 Market Risk, Ethics & Cautionary Tales	"flash crashes", "structure stress", "deepfakes", "fake posts", "stop-loss discipline", "enforcement outcomes"	• Structure Stress • Scams & Fraud • Consequences & Controls
T34 Promotion of Products, Events, and Community	"Telegram", "newsletters", "link in bio", "prompts/polls", "follows", "DMs", "votes", "funnels", "flywheel", "retention"	• Community Funnels • Engagement Prompts • Always-On CTAs

T35 Crypto: Between Institutionalization and Hype	"ETFs", "corporate treasuries", "payment integrations", "custody", "memecoins", "celebrity/influencer tokens", "pumps", "rugs", "lawsuits", "fraud risk"	• Institutional Rails • Speculative Waves • Risk Narratives
T36 Mindset, Identity, and the Trader Persona	"discipline", "patience", "conviction", "process > outcome", "persistence", "memes", "humor", "swagger", "belonging"	• Discipline Maxims • Journey Framing • Aspirational Swagger

Appendix B: Random Forest Results

Table B1 *Exhaustive random forest analysis results*

Target	Freq	ROC-AUC	AP; AP lift
1 - Women-First Money Empowerment (Financial Feminism)	n=10,492; Pos=194; Neg=10,298; Base=1.85%	0.649 ± 0.031	0.033 ± 0.004; 1.77
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	person	0.012	0.131; 0.407
2	suit	0.003	0.050; 0.090
3	trading chart	0.003	0.009; 0.115
4	car	0.000	-0.003; 0.082
5	boat	0.000	0.005; 0.074
6	tie	0.000	0.020; 0.064
7	banknote	0.000	0.003; 0.055
8	coin	-0.001	0.003; 0.050
9	money	-0.001	0.002; 0.064
Target	Freq	ROC-AUC	AP; AP lift
2 - Budgeting & Cashflow Systems	n=10,492; Pos=351; Neg=10,141; Base=3.35%	0.569 ± 0.022	0.042 ± 0.003; 1.25
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	person	0.005	0.057; 0.254
2	suit	0.001	0.026; 0.096
3	tie	0.001	0.021; 0.092
4	banknote	0.000	0.006; 0.060
5	boat	0.000	0.002; 0.105
6	car	0.000	0.005; 0.132
7	trading chart	-0.001	-0.004; 0.131
8	coin	-0.001	0.003; 0.065
9	money	-0.002	0.001; 0.064
Target	Freq	ROC-AUC	AP; AP lift
3 - Emergency Funds & HYSA (Liquidity)	n=10,492; Pos=250; Neg=10,242; Base=2.38%	0.518 ± 0.018	0.027 ± 0.001; 1.13
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	tie	0.002	0.021; 0.104
2	suit	0.001	0.010; 0.089
3	car	0.001	0.003; 0.144
4	trading chart	0.001	-0.006; 0.159
5	boat	0.001	0.007; 0.149
6	money	-0.000	0.004; 0.077
7	banknote	-0.000	0.006; 0.076
8	coin	-0.001	-0.008; 0.076
9	person	-0.001	-0.009; 0.125

Target	Freq	ROC-AUC	AP; AP lift
4 - Debt Reduction Systems	n=10,492; Pos=284; Neg=10,208; Base=2.71%	0.548 ± 0.053	0.032 ± 0.005; 1.18
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	trading chart	0.003	0.024; 0.168
2	car	0.002	0.020; 0.161
3	person	0.002	0.015; 0.134
4	tie	0.001	0.032; 0.117
5	suit	0.001	0.026; 0.116
6	coin	0.000	0.011; 0.084
7	money	-0.000	0.004; 0.066
8	banknote	-0.000	-0.001; 0.055
9	boat	-0.000	-0.000; 0.099
Target	Freq	ROC-AUC	AP; AP lift
5 - Credit Health & Cost of Capital	n=10,492; Pos=521; Neg=9,971; Base=4.97%	0.497 ± 0.037	0.051 ± 0.004; 1.03
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	person	0.001	0.007; 0.161
2	suit	0.001	0.005; 0.097
3	tie	-0.000	-0.005; 0.097
4	coin	-0.000	-0.002; 0.085
5	money	-0.001	-0.006; 0.069
6	boat	-0.001	-0.009; 0.101
7	banknote	-0.002	-0.009; 0.062
8	car	-0.002	-0.013; 0.154
9	trading chart	-0.003	-0.012; 0.174
Target	Freq	ROC-AUC	AP; AP lift
6 - Credit Cards, Rewards & Travel Hacking	n=10,492; Pos=240; Neg=10,252; Base=2.29%	0.536 ± 0.045	0.026 ± 0.003; 1.12
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	tie	0.001	0.029; 0.112
2	person	0.001	0.012; 0.138
3	suit	0.001	0.023; 0.116
4	car	0.000	0.011; 0.171
5	coin	0.000	0.004; 0.083
6	trading chart	0.000	0.004; 0.158
7	banknote	0.000	-0.000; 0.066
8	money	-0.000	-0.001; 0.064
9	boat	-0.001	-0.006; 0.093
Target	Freq	ROC-AUC	AP; AP lift
7 - Retirement & Account Hygiene	n=10,492; Pos=277; Neg=10,215; Base=2.64%	0.563 ± 0.027	0.031 ± 0.002; 1.16

Rank	Feature	Perm-AP	Perm-AUC; Gini
1	trading chart	0.002	0.028; 0.190
2	money	0.000	0.009; 0.166
3	suit	-0.000	0.008; 0.072
4	tie	-0.000	0.007; 0.080
5	banknote	-0.000	-0.002; 0.057
6	car	-0.000	0.000; 0.119
7	coin	-0.000	-0.001; 0.142
8	boat	-0.001	-0.002; 0.076
9	person	-0.002	-0.011; 0.097
Target	Freq	ROC-AUC	AP; AP lift
8 - Investing 101 & Long-Termism	n=10,492; Pos=1,435; Neg=9,057; Base=13.68%	0.586 ± 0.022	0.170 ± 0.012; 1.25
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	tie	0.036	0.087; 0.267
2	person	0.015	0.033; 0.185
3	suit	0.012	0.016; 0.150
4	car	0.003	0.007; 0.097
5	boat	0.002	0.002; 0.072
6	trading chart	0.002	0.002; 0.095
7	coin	0.000	0.002; 0.051
8	banknote	0.000	0.001; 0.045
9	money	-0.001	0.001; 0.037
Target	Freq	ROC-AUC	AP; AP lift
9 - Taxes & Paycheck Optimization	n=10,492; Pos=1,205; Neg=9,287; Base=11.48%	0.540 ± 0.010	0.130 ± 0.005; 1.13
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	person	0.010	0.038; 0.290
2	tie	0.005	0.009; 0.085
3	banknote	0.002	-0.001; 0.068
4	suit	0.001	0.002; 0.079
5	trading chart	-0.000	-0.000; 0.142
6	money	-0.001	-0.001; 0.060
7	boat	-0.001	-0.003; 0.097
8	car	-0.002	0.000; 0.117
9	coin	-0.003	-0.006; 0.062
Target	Freq	ROC-AUC	AP; AP lift
10 - Insurance & Protection + Estate Basics	n=10,492; Pos=486; Neg=10,006; Base=4.63%	0.588 ± 0.012	0.058 ± 0.002; 1.26
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	trading chart	0.006	0.039; 0.202
2	car	0.003	0.016; 0.120
3	tie	0.002	0.036; 0.085

4	person	0.002	0.011; 0.095
5	boat	0.002	0.009; 0.090
6	suit	0.002	0.027; 0.074
7	coin	0.001	0.012; 0.169
8	money	0.000	0.006; 0.117
9	banknote	-0.000	-0.001; 0.047
Target	Freq	ROC-AUC	AP; AP lift
11 - Housing & Real Estate Literacy	n=10,492; Pos=841; Neg=9,651; Base=8.02%	0.551 ± 0.018	0.097 ± 0.004; 1.21
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	person	0.010	0.035; 0.225
2	suit	0.010	0.019; 0.098
3	tie	0.008	0.012; 0.093
4	trading chart	0.005	0.013; 0.155
5	car	0.002	-0.001; 0.129
6	boat	0.001	0.001; 0.115
7	money	-0.000	0.000; 0.067
8	banknote	-0.001	-0.003; 0.052
9	coin	-0.001	0.001; 0.067
Target	Freq	ROC-AUC	AP; AP lift
12 - Income Growth: Career Capital	n=10,492; Pos=294; Neg=10,198; Base=2.80%	0.675 ± 0.009	0.055 ± 0.003; 1.96
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	suit	0.017	0.155; 0.289
2	tie	0.014	0.122; 0.204
3	trading chart	0.005	0.026; 0.127
4	person	0.002	0.023; 0.107
5	coin	0.000	0.001; 0.072
6	boat	-0.000	0.004; 0.052
7	car	-0.000	0.006; 0.076
8	banknote	-0.001	-0.004; 0.022
9	money	-0.001	-0.006; 0.052
Target	Freq	ROC-AUC	AP; AP lift
13 - Side Hustles & Entrepreneurship	n=10,492; Pos=765; Neg=9,727; Base=7.29%	0.708 ± 0.025	0.214 ± 0.035; 2.93
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	suit	0.129	0.138; 0.207
2	tie	0.123	0.131; 0.176
3	car	0.042	0.049; 0.228
4	person	0.038	0.071; 0.100
5	coin	0.012	0.020; 0.091
6	trading chart	0.011	0.009; 0.065
7	banknote	0.004	0.007; 0.036

8	boat	0.002	0.003; 0.050
9	money	-0.000	0.001; 0.047
Target	Freq	ROC-AUC	AP; AP lift
14 - Creator Economy & Audience Growth	n=10,492; Pos=246; Neg=10,246; Base=2.34%	0.480 ± 0.020	0.023 ± 0.002; 0.99
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	car	0.001	0.012; 0.190
2	money	0.000	-0.004; 0.086
3	trading chart	-0.000	-0.008; 0.159
4	banknote	-0.000	-0.003; 0.065
5	coin	-0.000	-0.004; 0.103
6	boat	-0.001	-0.013; 0.094
7	suit	-0.001	-0.013; 0.098
8	tie	-0.001	-0.018; 0.086
9	person	-0.002	-0.025; 0.120
Target	Freq	ROC-AUC	AP; AP lift
15 - CTAs, Lead Magnets & Community Offers	n=10,492; Pos=310; Neg=10,182; Base=2.95%	0.502 ± 0.043	0.030 ± 0.003; 1.02
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	banknote	0.001	0.004; 0.064
2	trading chart	0.001	0.013; 0.181
3	tie	0.000	0.004; 0.107
4	suit	-0.000	0.002; 0.105
5	coin	-0.000	-0.002; 0.085
6	money	-0.001	-0.014; 0.074
7	boat	-0.001	-0.003; 0.111
8	car	-0.002	-0.020; 0.144
9	person	-0.003	-0.021; 0.129
Target	Freq	ROC-AUC	AP; AP lift
16 - Advocacy & Civic Issues via Money	n=10,492; Pos=1,537; Neg=8,955; Base=14.65%	0.610 ± 0.010	0.192 ± 0.005; 1.31
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	tie	0.048	0.101; 0.313
2	suit	0.015	0.019; 0.270
3	person	0.011	0.022; 0.117
4	coin	0.006	0.009; 0.058
5	car	0.004	0.006; 0.071
6	trading chart	0.003	0.003; 0.068
7	boat	0.002	0.001; 0.040
8	banknote	0.001	0.003; 0.031
9	money	-0.000	0.003; 0.034
Target	Freq	ROC-AUC	AP; AP lift

17 - Money & Relationships	**n=10,492; Pos=270; Neg=10,222; Base=2.57%**	**0.568 ± 0.029**	**0.033 ± 0.004; 1.29**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**suit**	**0.006**	**0.042; 0.138**
2	**tie**	**0.004**	**0.019; 0.108**
3	**trading chart**	**0.002**	**0.022; 0.167**
4	**car**	**0.002**	**0.017; 0.166**
5	**person**	**0.001**	**0.019; 0.147**
6	**boat**	**-0.001**	**-0.000; 0.078**
7	**banknote**	**-0.001**	**-0.004; 0.056**
8	**coin**	**-0.002**	**-0.007; 0.075**
9	**money**	**-0.002**	**-0.012; 0.065**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
18 - Parenting, Kids & Intergenerational Planning	**n=10,492; Pos=248; Neg=10,244; Base=2.36%**	**0.547 ± 0.020**	**0.028 ± 0.002; 1.17**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**car**	**0.001**	**0.017; 0.145**
2	**suit**	**0.001**	**0.009; 0.090**
3	**money**	**0.001**	**0.013; 0.091**
4	**person**	**0.001**	**0.012; 0.136**
5	**boat**	**0.000**	**0.004; 0.111**
6	**banknote**	**0.000**	**0.002; 0.077**
7	**coin**	**-0.000**	**-0.000; 0.087**
8	**tie**	**-0.000**	**-0.002; 0.076**
9	**trading chart**	**-0.000**	**0.006; 0.187**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
19 - Values, Mindset & Mental Health	**n=10,492; Pos=295; Neg=10,197; Base=2.81%**	**0.510 ± 0.019**	**0.029 ± 0.002; 1.03**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**tie**	**0.001**	**0.023; 0.136**
2	**suit**	**0.001**	**0.019; 0.126**
3	**money**	**0.000**	**0.007; 0.070**
4	**coin**	**-0.000**	**0.002; 0.071**
5	**banknote**	**-0.000**	**0.004; 0.067**
6	**boat**	**-0.000**	**0.000; 0.096**
7	**car**	**-0.001**	**-0.003; 0.150**
8	**trading chart**	**-0.002**	**-0.018; 0.149**
9	**person**	**-0.002**	**-0.007; 0.136**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
20 - Consumer Savviness & Frugality	**n=10,492; Pos=507; Neg=9,985; Base=4.83%**	**0.501 ± 0.009**	**0.049 ± 0.001; 1.01**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**

1	boat	0.000	-0.001; 0.144
2	banknote	-0.000	-0.002; 0.058
3	money	-0.000	0.001; 0.057
4	tie	-0.000	-0.004; 0.141
5	coin	-0.001	0.003; 0.070
6	suit	-0.001	-0.013; 0.138
7	car	-0.002	-0.012; 0.148
8	person	-0.003	-0.010; 0.106
9	trading chart	-0.003	-0.012; 0.139
Target	Freq	ROC-AUC	AP; AP lift
21 - Safety, Fraud & OPSEC	n=10,492; Pos=876; Neg=9,616; Base=8.35%	0.561 ± 0.019	0.099 ± 0.006; 1.18
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	trading chart	0.007	0.031; 0.209
2	coin	0.003	0.015; 0.174
3	car	0.003	0.007; 0.152
4	tie	0.002	0.013; 0.084
5	boat	0.000	0.001; 0.087
6	suit	0.000	0.007; 0.074
7	person	0.000	-0.003; 0.072
8	banknote	-0.000	-0.003; 0.065
9	money	-0.001	-0.000; 0.084
Target	Freq	ROC-AUC	AP; AP lift
22 - Small-Business Finance & Cost Structure	n=10,492; Pos=249; Neg=10,243; Base=2.37%	0.547 ± 0.035	0.029 ± 0.004; 1.24
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	boat	0.002	0.017; 0.117
2	person	0.002	0.012; 0.145
3	trading chart	0.001	0.015; 0.192
4	coin	0.000	0.003; 0.084
5	banknote	0.000	0.000; 0.058
6	tie	-0.000	0.012; 0.102
7	suit	-0.000	0.009; 0.101
8	car	-0.000	-0.007; 0.145
9	money	-0.001	-0.002; 0.056
Target	Freq	ROC-AUC	AP; AP lift
23 - Women of Color / First-Gen Wealth Narratives	n=10,492; Pos=253; Neg=10,239; Base=2.41%	0.555 ± 0.018	0.029 ± 0.001; 1.21
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	person	0.002	0.027; 0.149
2	trading chart	0.001	0.007; 0.158
3	tie	0.001	0.018; 0.091
4	coin	0.000	0.006; 0.144

5	**suit**	**0.000**	**0.008; 0.090**
6	**boat**	**0.000**	**0.003; 0.093**
7	**car**	**-0.000**	**0.004; 0.147**
8	**banknote**	**-0.002**	**-0.003; 0.049**
9	**money**	**-0.002**	**-0.006; 0.079**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
24 - Health × Money	**n=10,492; Pos=559; Neg=9,933; Base=5.33%**	**0.536 ± 0.031**	**0.060 ± 0.005; 1.13**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**person**	**0.003**	**0.028; 0.186**
2	**tie**	**0.002**	**0.019; 0.118**
3	**coin**	**0.001**	**0.002; 0.082**
4	**banknote**	**0.000**	**0.002; 0.071**
5	**boat**	**0.000**	**-0.003; 0.114**
6	**suit**	**-0.000**	**-0.002; 0.082**
7	**money**	**-0.000**	**-0.000; 0.062**
8	**trading chart**	**-0.002**	**-0.009; 0.135**
9	**car**	**-0.002**	**0.001; 0.150**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
25 - Housing Access & Affordability	**n=10,492; Pos=270; Neg=10,222; Base=2.57%**	**0.526 ± 0.014**	**0.027 ± 0.001; 1.06**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**boat**	**-0.000**	**0.003; 0.149**
2	**car**	**-0.000**	**0.005; 0.143**
3	**person**	**-0.001**	**0.004; 0.174**
4	**money**	**-0.001**	**-0.006; 0.074**
5	**coin**	**-0.001**	**-0.002; 0.108**
6	**suit**	**-0.001**	**-0.011; 0.071**
7	**banknote**	**-0.001**	**-0.005; 0.068**
8	**tie**	**-0.001**	**-0.012; 0.061**
9	**trading chart**	**-0.002**	**0.009; 0.152**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
26 - Market Contexts & Economic Narratives	**n=10,492; Pos=452; Neg=10,040; Base=4.31%**	**0.570 ± 0.015**	**0.052 ± 0.002; 1.20**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**trading chart**	**0.005**	**0.043; 0.272**
2	**banknote**	**0.002**	**0.009; 0.056**
3	**boat**	**0.001**	**0.008; 0.099**
4	**suit**	**0.000**	**0.010; 0.083**
5	**car**	**0.000**	**0.003; 0.136**
6	**coin**	**0.000**	**0.005; 0.091**
7	**tie**	**0.000**	**0.010; 0.077**
8	**person**	**-0.001**	**-0.002; 0.128**

9	money	-0.001	-0.002; 0.059
Target	Freq	ROC-AUC	AP; AP lift
27 - Tech Tools & Financial Apps	n=10,492; Pos=1,057; Neg=9,435; Base=10.07%	0.600 ± 0.018	0.145 ± 0.014; 1.43
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	suit	0.035	0.056; 0.154
2	tie	0.026	0.038; 0.145
3	person	0.020	0.058; 0.241
4	trading chart	0.007	0.013; 0.128
5	car	0.004	0.006; 0.097
6	coin	0.002	0.012; 0.075
7	boat	-0.000	-0.000; 0.072
8	banknote	-0.001	0.001; 0.040
9	money	-0.003	0.002; 0.048
Target	Freq	ROC-AUC	AP; AP lift
28 - Sustainability & Ethical Finance	n=10,492; Pos=565; Neg=9,927; Base=5.39%	0.559 ± 0.033	0.074 ± 0.007; 1.37
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	suit	0.011	0.027; 0.111
2	car	0.010	0.007; 0.147
3	tie	0.009	0.010; 0.096
4	person	0.008	0.023; 0.165
5	coin	0.005	0.006; 0.077
6	boat	0.005	0.008; 0.144
7	money	0.003	0.003; 0.055
8	banknote	0.002	0.005; 0.069
9	trading chart	-0.000	0.001; 0.136
Target	Freq	ROC-AUC	AP; AP lift
29 - Behavioral Economics & Nudges	n=10,492; Pos=371; Neg=10,121; Base=3.54%	0.581 ± 0.019	0.044 ± 0.002; 1.24
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	trading chart	0.003	0.030; 0.168
2	car	0.003	0.031; 0.187
3	tie	0.002	0.043; 0.091
4	suit	0.002	0.035; 0.099
5	coin	0.002	0.020; 0.126
6	person	0.000	0.012; 0.126
7	banknote	0.000	0.003; 0.058
8	boat	-0.000	0.005; 0.086
9	money	-0.001	0.004; 0.059
Target	Freq	ROC-AUC	AP; AP lift

30 - Faith, Purpose & Money Philosophy	**n=10,492; Pos=301; Neg=10,191; Base=2.87%**	**0.528 ± 0.025**	**0.033 ± 0.003; 1.16**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**tie**	**0.002**	**0.017; 0.096**
2	**boat**	**0.002**	**0.010; 0.136**
3	**suit**	**0.002**	**0.015; 0.092**
4	**car**	**0.001**	**0.012; 0.155**
5	**trading chart**	**0.000**	**0.004; 0.169**
6	**person**	**0.000**	**0.017; 0.114**
7	**banknote**	**-0.000**	**0.001; 0.069**
8	**money**	**-0.001**	**0.005; 0.084**
9	**coin**	**-0.001**	**0.004; 0.085**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
31 - Trader Education & Execution (core pedagogy)	**n=10,492; Pos=961; Neg=9,531; Base=9.16%**	**0.821 ± 0.010**	**0.367 ± 0.019; 4.01**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**trading chart**	**0.244**	**0.216; 0.820**
2	**coin**	**0.045**	**0.017; 0.062**
3	**car**	**0.010**	**0.005; 0.017**
4	**boat**	**0.002**	**0.002; 0.011**
5	**banknote**	**0.001**	**0.003; 0.011**
6	**person**	**-0.002**	**0.004; 0.031**
7	**tie**	**-0.005**	**0.002; 0.010**
8	**money**	**-0.005**	**0.005; 0.030**
9	**suit**	**-0.006**	**0.002; 0.008**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
32 - Fintech Infrastructure & Market Access	**n=10,492; Pos=220; Neg=10,272; Base=2.10%**	**0.656 ± 0.039**	**0.038 ± 0.007; 1.83**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**
1	**trading chart**	**0.003**	**0.031; 0.225**
2	**suit**	**0.003**	**0.059; 0.150**
3	**tie**	**0.002**	**0.023; 0.120**
4	**coin**	**0.001**	**0.027; 0.121**
5	**person**	**0.001**	**0.037; 0.128**
6	**car**	**0.000**	**0.004; 0.086**
7	**boat**	**-0.000**	**0.007; 0.066**
8	**banknote**	**-0.001**	**-0.001; 0.048**
9	**money**	**-0.002**	**0.001; 0.055**
Target	**Freq**	**ROC-AUC**	**AP; AP lift**
33 - Market Risk, Ethics & Cautionary Tales	**n=10,492; Pos=656; Neg=9,836; Base=6.25%**	**0.570 ± 0.023**	**0.076 ± 0.006; 1.22**
Rank	**Feature**	**Perm-AP**	**Perm-AUC; Gini**

1	suit	0.006	0.042; 0.141
2	trading chart	0.005	0.019; 0.172
3	coin	0.005	0.026; 0.132
4	person	0.002	0.003; 0.095
5	boat	0.001	0.001; 0.090
6	tie	0.000	0.005; 0.077
7	banknote	-0.000	0.005; 0.089
8	money	-0.001	0.002; 0.073
9	car	-0.002	0.004; 0.131
Target	Freq	ROC-AUC	AP; AP lift
34 - Promotion of Products, Events, and Community	n=10,492; Pos=250; Neg=10,242; Base=2.38%	0.656 ± 0.059	0.070 ± 0.022; 2.92
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	trading chart	0.029	0.081; 0.309
2	coin	0.025	0.064; 0.211
3	money	0.010	-0.002; 0.092
4	car	0.010	0.004; 0.082
5	suit	0.004	0.015; 0.066
6	person	0.004	-0.008; 0.069
7	banknote	0.001	-0.004; 0.038
8	tie	0.000	0.024; 0.073
9	boat	-0.002	-0.009; 0.060
Target	Freq	ROC-AUC	AP; AP lift
35 - Crypto: Between Institutionalization and Hype	n=10,492; Pos=814; Neg=9,678; Base=7.76%	0.857 ± 0.018	0.445 ± 0.038; 5.74
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	coin	0.262	0.140; 0.513
2	money	0.056	0.012; 0.294
3	tie	0.051	0.030; 0.032
4	suit	0.048	0.029; 0.028
5	car	0.036	0.016; 0.053
6	trading chart	0.030	0.016; 0.027
7	person	0.007	0.011; 0.018
8	banknote	0.006	0.005; 0.026
9	boat	-0.001	0.003; 0.008
Target	Freq	ROC-AUC	AP; AP lift
36 - Mindset, Identity, and the Trader Persona	n=10,492; Pos=1,064; Neg=9,428; Base=10.14%	0.721 ± 0.012	0.285 ± 0.023; 2.81
Rank	Feature	Perm-AP	Perm-AUC; Gini
1	boat	0.090	0.038; 0.242
2	suit	0.077	0.070; 0.114
3	tie	0.065	0.067; 0.101
4	trading chart	0.058	0.039; 0.140

5	**car**	**0.052**	**0.029; 0.137**
6	**coin**	**0.016**	**0.027; 0.096**
7	**person**	**0.015**	**0.028; 0.081**
8	**banknote**	**0.002**	**0.009; 0.049**
9	**money**	**0.000**	**0.005; 0.041**