SustDict: a customized lexical-based dictionary for CSR

Emma Zavarrone[1], Alessia Forciniti[2], Marta Muscariello[3]

[1] IULM University – emma.zavarrone@iulm.it

[2] IULM University – alessia.forciniti@iulm.it

[3] IULM University – marta.muscariello@iulm.it

Abstract

In 2001, EU defined Corporate Social Responsibility (CSR) as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis. Being socially responsible means not only fulfilling legal expectations, but also going beyond compliance and investing more into human capital, the environment, and the relations with stakeholders". Following this definition, the CSR' pillars are represented by environmental, social, and economic sustainability, and must be communicated to the society through appropriate reports. Sentiment Analysis (SA) can capture the tone and opinion of reports by detecting polarity and identifying the class of emotion hidden in the documents. SA can be conducted at different levels: document, sentence and aspect. However, SA depends on the availability of dictionary. This dictionary has a great influence on results and the corresponding analyses. The construction of a universal sentiment dictionary that fits all domains is very complex because word sentiment depends on the domain it has been used for. The dictionaries of given domain cannot easily be applied to other domains because terms can have opposite sentiments when used in different situations. For this reason, we developed a customized lexical-based dictionary for Italian sustainability terms using methods for sentiment polarity categorization based on polarity features, such as adjectives, verbs, and nouns, by combining textual analysis techniques with the properties of social network analysis. The corpus is composed by environmental, social, and governance (ESG) disclosure strategies for Italian listed companies that closed the financial year at 31st December 2021.

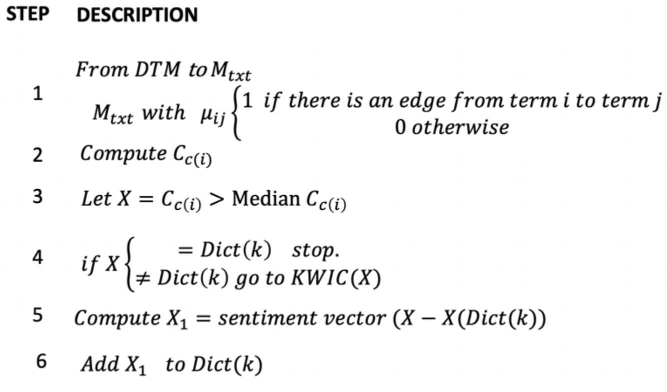
Keywords: sentiment analysis, lexicon-based methods, customized-lexicon, polarity classification, performance measures

1. Introduction

The awareness of the paradigm of sustainability (Lukman & Glavic, 2007) has enriched and diversified many fields of study, including the corporate organizational culture (Aguinis & Glavas, 2012). According to the European Commission (2011), the role of business concerning the sustainable development is related to activities based on corporate social responsibility (CSR) and thus aimed at reducing the impact of the negative effects of business (Carpenter & White, 2004). At the same time, the sustainable development represents a model of strategic management to gain competitive advantage bringing the CSR to the level of a system of "governance" of transactions and relations between the company and its stakeholders, maximizing long-term profits (Chandler & Werther, 2014). The widest literature is based on the relationships between CSR and corporate performance (McWilliams & Siegel, 2011), while less developed is the integrated view of corporate sustainability management and CSR (Aguinis & Glavas, 2012). In this perspective, we are not able to clearly determine a well-defined language used by companies to communicate the CSR to stakeholders. One of the methods used to study communication and obtain an inclusive community perspective on a specific area is Sentiment Analysis (SA). In order to know the opinion on stock markets (e.g. Arnold & Vrugt, 2008), economic systems policies, SA finds application in fields such as business, organization, government. SA is performed by means of machine learning or lexicon-based methods or the combination of these two in a hybrid approach. In this paper, our interest is on the lexicon-based (LB) approaches, which according to some scenarios (e.g. Vassallo *et al.,* 2020) are preferred to supervised classification methods, since they are more robust for the classification across domains and present a greater scalability. LB methods allow the interpretation of a single word on its context-dependency and are, essentially, related to polarity lexicons which attribute a valence of positive, negative or neutral sentiment to a text (e.g. -1, 0, 1). However, the classification of polarity may depend on the domain in which each word fits and the lexicons of reference cannot cover all the meanings of the words specific to a domain. The field of application of SA offers different lexicons, some of which are based on general knowledge of sentiment by including different areas such as sports, politics, language of social media (Nielsen, 2011). Other lexicons are domain-based in sector specificities such as the economic-financial environment (Henry, 2008; Loughran & Mcdonald, 2011). If on the one hand the lexicons of the sentiment of general knowledge have the advantage of a wider coverage of the phenomena, on the other hand, they have lower performance than the domain-based lexicons, since some terms may neither be present in a general lexicon nor can have a different contextual classification. Technically, customized lexicon requires a manual labelling data often very expensive. Thus, pre-constituted dictionaries specific to a language and a domain of interest are increasing, as well as widespread used. These dictionaries are useful in contexts where the available data are reduced. This perspective is very common for low-resource languages such as Italian. Different general lexicons of sentiment have been proposed for Italian, they are often the results of translation from English. Among the most popular, we can mention SentiWordNet (Esuli & Sebastiani, 2006) and the National Research Council (NRC) Emotion lexicon (Mohammad & Turney, 2010). The same limitations are connected to domain-based lexicons mainly developed in English (cf. Henry, 2008; Loughran & Mcdonald, 2011) but, with reference to sustainability management there is a lack of resources both in English and other languages. Thus, our contribution aims to select an effective method for approaching SA on CSR in Italian. The paper consists in a practical evaluation of the performance of LB methods by means of two steps: the first step is oriented to assess the use of general lexicon, and the second one is focused on the evaluation of a customized-lexicon given by the implementation of general lexicons with lexical items connected to CSR. The lexical specificities are detected combining textual and Social Network Analyses and specifically we compare SentiWordNet and NRC Lexicon.

2. Method

The method under development goes into the direction of proposing an innovative model to improve the lexicon performance and implement the lexical items related to CSR in Italian corporate communication. Our innovative methodology is based on the building of a customized-lexicon by means of a multi-stage model that combines text analysis with Social Network Analysis (SNA). The main advantage of our model is the possibility to detect unigrams and bigrams which characterize a given corpus and to attribute them a positive or negative valence performing a syntactic and semantic analysis of each lexical item selected. For this purpose, we used a cross validation procedure based on the sub-setting the corpus in five sub-corpora, each one composed by a random sample of 26 CSR reports. We develop the model (*Figure 1*) in sample 1 (train set) and we use the rest of samples to test the method (test set). The stages of the method are: 1) corpus creation and a document-term matrix $\mathbf{DTM}_{dxt}$, with *d* CSR reports, and *t* terms, construction. The data pre-processing operations have been implemented through typical operations of removal of punctuation, numbers, stop words and company's name, and lemmatization. Specifically, we built the **DTM** normalizing the schemes for weighting from tf (term-frequency) to tf-IDF (term frequency-Inverse Document Frequency) so as to quantify the specificity of a term in inverse proportion to the number of documents in which it occurs; 2) Transformation **DTM** in an adjacency matrix, $\mathbf{M}_{txt}$, based on the study of co-occurrences which allows to keep mark of the structure of each text without breaking the links with semantic analysis. We take advantage of social network features to detect unigrams and bigrams using the centrality measure known as Freeman's closeness C_{ci}, (Freeman, 1978) which represents the shortest path from one to the other; 3) All terms greater to the median value of closeness represent the words to implement into lexicon; 4) We compare the C_{Ci} selected terms with lexicon terms and we implement a decision rule based on the presence/absence paradigm. If the C_{Ci} terms already do exist then the procedure will be stopped, otherwise the return to the syntactic and semantic analysis that precedes the pre-treatment of the text will be realized; 5) By means of analysis both of collocations among the terms and the context (*kwic* - keyword in the context), we assign a sentiment label. The new sentiment (positive or negative) will be adjunct to the *k'*. The k' will be tested in four samples (i.e. testing set). We applied performance measures for binary sentiment classification to validate our method. More precisely, we estimated the overall effectiveness of the classifier in terms of *accuracy*; the effectiveness of the classifier to identify positive labels as *sensitivity* or *recall;* the effectiveness of the classifier to determine the negative labels in terms of *specificity*.



STEP	DESCRIPTION
1	$From\ DTM\ to\ M_{txt}$ $M_{txt}\ with\ \ \mu_{ij}\begin{cases}1\ if\ there\ is\ an\ edge\ from\ term\ i\ to\ term\ j\\ 0\ otherwise\end{cases}$
2	$Compute\ C_{c(i)}$
3	$Let\ X = C_{c(i)} >$ Median $C_{c(i)}$
4	$if\ X\begin{cases}= Dict(k)\quad stop.\\ \neq Dict(k)\ go\ to\ KWIC(X)\end{cases}$
5	$Compute\ X_1 = sentiment\ vector\ (X - X(Dict(k)))$
6	$Add\ X_1\ \ to\ Dict(k)$

Figure 1. Algorithm structure: steps

3. Data and results

We collected the sustainability reports drafted by the complete listing of Italian companies listed on stock exchanges that closed the financial year at 31st December 2021, in accordance with the Legislative Decree No. 254 of 30th December 2016[1]. Of the complete list that includes 196 companies, we have selected 130 documents, because of the unavailability of some sources or the drafting of the report in English. At first, we create the corpus and applied the data treatment procedure in order to create a vocabulary. After sub-setting the corpus, we analysed the vocabulary of each sample through an exploratory textual data analysis (ETDA (*Table 1*). We observed – for each sample – a generalist vocabulary with few references to a specialized language or related to technicalities of the domain. To apply our model and to verify the effectiveness of the customized-lexicon on CSR, we propose the comparison of performance of two of the most popular general lexicons (indicated by *Dict(k)*) for Italian before and after their implementation. The first lexicon used is SentiWordNet *(k=1)* developed by Esuli & Sebastiani (2006) from a corpus-based semantic approach to detect positive, negative and objective (neutral) polarity. The second lexicon used is NRC *(k=2)* Emotion Lexicon developed by Mohammad and Turney in 2010. NRC lexicon is used to detect simultaneously the semantic orientation (positive or negative) and the emotions. Mohammad and Turney used the Mechanical Turk (Amazon Service) obtaining human annotations at sense level rather than at word level in an efficient way. In SentiWordNet scores are measured on a continuous scale with different intensity from 0 to 1. NRC Emotion Lexicon presents instead a classification based on binary polarity negative or positive (-1 or 1). Thus, to perform the comparison between two lexicons, we converted each score of SentiWordNet into a binary sentiment category labelled as negative or positive, by using only the sign of the score and removing the polarity equal to 0 marked as objective or neutral.

Table. 1. ETDA

Sample	*N. lemmas*	*Unigrams*	*Bigrams*
First	21,754	gestione, gri, rischio, emissione, …	consumo_energetico, capitale_umano, codice_etico, lotta_corruzione, …
Second	21,867	salute, consumo, impianto, sicurezza, …	aspetto_materiale, gestione_rischio, pari_opportunità,corporate_governance, …
Third	22,496	modello, capitale, formazione, obiettivo, …	controllo_rischio, collegio_sindacale, risorsa_umana, ...
Fourth	20,915	mercato, salute, impatto, gri, …	gestione_responsabile, ricerca_sviluppo, fonte_rinnovabile, donna_uomo, …
Fifth	20,644	ambientale, sviluppo, processo, sistema, …	salute_sicurezza, controllo_interno, materia_prima, etica_integrità, …

The overall sentiment of 130 sustainability reports showed a predominance of positive sentiment both by means of SentiWordNet and NRC Emotion Lexicon. More precisely, we detected 257,207 negative words and 508,598 positive ones by using the SentiWordNet lexicon; while we observed 120,905 negative words and 397,615 positive ones with NRC. We obtained the same prevalence of positive sentiment for each sample, varying equally between the two

[1] https://www.borsaitaliana.it/notizie/sotto-la-lente/informazioni-non-finanziarie.htm

lexicons from 19% to 22%. The negative polarity is also almost distributed in a similar way between the two lexicons varying from 18% to 23%.

Table 2. Performance measures of the samples before the implementation

	SentiWordNet			*NRC*		
Sample	*Accuracy*	*Recall*	*Specificity*	*Accuracy*	*Recall*	*Specificity*
First	0.53	0.45	0.69	0.68	0.30	0.78
Second	0.53	0.38	0.69	0.64	0.28	0.77
Third	0.51	0.37	0.68	0.53	0.28	0.80
Fourth	0.55	0.36	0.69	0.66	0.27	0.79
Fifth	0.65	0.36	0.70	0.52	0.26	0.79

The performance measures computed on both lexicons before the implementation (*Table 2*) show that NRC lexicon detected on average a better accuracy in the sentiment classification. Its average accuracy is greater of 5.2% than SentiWordNet's performance. This depends on its specificity, that is NRC's ability to better detect the words with negative polarity in this context for 78.6%. On the contrary, SentiWordNet well performed on the positive side, showing a greater average percentage of recall equal to 38.4%. Our model allowed to detect 65 unigrams and 82 bigrams greater to the median of closeness. By the annotation procedure aimed to attribute positive or negative valence to each item, we determined 57 positive unigrams and 8 negative ones, 61 positive bigrams and 21 negative ones. The implementation of the lexicons through the new 147 lexical polarized items has determined that the customized-lexicon of SentiWordNet is composed of 20,240 items and NRC Lexicon consists of 5,615 items. The performance measures on the implemented lexicons showed an overall improvement (*Table 3*) which on average varies between the different measures from 0.8% to 8.8%. The *Table* 3 confirmed a 2% higher mean accuracy for NRC lexicon. However, despite the greater effectiveness of the NRC classifier in identifying negative labels, the implementation of the lexicon has mostly improved the SentiwordNet's specificity by 0.6% more than the NRC. The domain-based implementation favoured the performance of SentiWordNet for 5.2% compared to 4.6% of NRC. The performance in the classification of positive labels has instead confirmed the greater effectiveness of SentiWordNet as recall, since NRC increased on average its recall by only 0.8% compared to 2.8% of SentiWordNet.

Table 3. Performance measures of the samples after the implementation

	SentiWordNet			*NRC*		
Sample	*Accuracy*	*Recall*	*Specificity*	*Accuracy*	*Recall*	*Specificity*
First (baseline)	0.59	0.52	0.76	0.69	0.32	0.84
Second	0.60	0.39	0.74	0.71	0.28	0.83
Third	0.57	0.39	0.69	0.70	0.28	0.85
Fourth	0.68	0.39	0.77	0.68	0.28	0.81
Fifth	0.68	0.39	0.75	0.69	0.27	0.83

4. Conclusions and future work

We studied the paradigm of sustainability in the Italian corporate culture, by analysing the sentiment of the sustainability reports drafted by listed companies. We proposed an innovative method for developing a customized lexical-based dictionary able to investigate the sustainability in Italian language. By integrating bag of words coding, linguistic approach and SNA, we implemented and compared two Italian general lexicons. To test the efficacy of our model, we split under cross validation perspective our corpus in five random samples: one used such as train set and four as test set. The procedure highlights a significant improvement of performance measurements in all samples compared with the baseline sample. Future developments lead us to include other sources to improve data training and apply a neural network approach. The development of SustDict package, based on the algorithm structure proposed, represents a further step to deal with.

References

Aguinis H, Glavas A. (2012). What we know and don’t know about corporate social responsibility: A review and research agenda. *Journal of Management,* vol. (38): 932-968.

Arnold, I., Vrugt, E. (2008). Fundamental uncertainty and stock market volatility. *Applied Financial Economics*, vol. (18): 1425-1440.

Carpenter, G. & White, P. (2004). Sustainable development: Finding the real business case. *Corporate Environmental Strategy. International Journal for Sustainable Business*, vol. (11): 51-56.

Chandler, D. & Werther, W.B. (2014). *Strategic Corporate Social Responsibility. Stakeholders, Globalization, and Sustainable Value Creation*. 3^{a} ed. SAGE Publications.

Esuli, A. & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *In Proceedings of LREC 2006.*

European Commission (2011). A renewed EU strategy 2011-14 for Corporate Social Responsibility.

Freeman, L.C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks*, vol. (1), pp.215-239.

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written?. *Journal of Business Communication* (45): 363-407.

Loughran, T. & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance (66.1): pp. 35-65.

Lukman, R. & Glavic, P. (2007). Review of sustainability terms and their definitions. *Journal of Cleaner Production*, vol. (15): 1875-1885.

McWilliams, A. & Siegel, D.S. (2011). Creating and capturing value: Strategic corporate social responsibilty, resource-based theory, and sustainable competitive advantage. *Journal of Management*, vol. (37): 1480-1495.

Mohammad, S. & Turney, P. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*

Nielsen, F.A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *In Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages*, pp. 93-98.

Vassallo, M., *et al.* (2020). Polarity Imbalance in Lexicon-based Sentiment Analysis. *In Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020.* Accademia University Press.