Highlights

A Hybrid Lexicon-based and Neural Approach for eXplainable Polarity Detection

Marco Polignano,Valerio Basile,Pierpaolo Basile,Giuliano Gabrieli,Marco Vassallo,Cristina Bosco

- Definition of an affective lexicon for the Italian language, WMAL
- Definition of a hybrid lexicon-deep learning classification model
- Definition of an explanation strategy for justifying the classifications obtained

A Hybrid Lexicon-based and Neural Approach for eXplainable Polarity Detection

Marco Polignano[a,*] (Assistant Professor), Valerio Basile[b] (Assistant Professor), Pierpaolo Basile[a] (Associate Professor), Giuliano Gabrieli[c], Marco Vassallo[c] and Cristina Bosco[b] (Assistant Professor)

[a] *University of Bari Aldo Moro, Via E. Orabona 4, Bari, 70125, Bari, Apulia, Italy*
[b] *University of Turin, Via Giuseppe Verdi 8, Torino, 10124, Torino, Piemonte, Italy*
[c] *CREA - Research Centre for Agricultural Policies and Bio-economy, Italy*

ARTICLE INFO

Keywords:
sentiment analysis
polarity detection
lexicon
WMAL
BERT
explanation
deep learning
machine learning

ABSTRACT

In this work, we propose BERT-WMAL, a hybrid model that brings together information coming from data through the recent transformer deep learning model and those obtained from a polarized lexicon. The result is a model for sentence polarity that manages to have performances comparable with those at the state-of-the-art, but with the advantage of being able to provide the end-user with an explanation regarding the most important terms involved with the provided prediction. The model has been evaluated on three polarity detection Italian dataset, i.e., SENTIPOLC, AGRITREND and ABSITA. While the first contains 7,410 tweets released for training and 2,000 for testing, the second and the third respectively include 1,000 tweets without splitting, and 2,365 reviews for training, 1,171 for testing. The use of lexicon-based information proves to be effective in terms of the F1 measure since it shows an improvement of F1 score on all the observed dataset: from 0.664 to 0.669 (i.e, 0.772%) on AGRITREND, from 0.728 to 0.734 (i.e., 0.854%) on SENTIPOLC and from 0.904 to 0.921 (i.e, 1.873%) on ABSITA. The usefulness of this model not only depends on its effectiveness in terms of the F1 measure, but also on its ability to generate predictions that are more explainable and especially convincing for the end-users. We evaluated this aspect through a user study involving four native Italian speakers, each evaluating 64 sentences with associated explanations. The results demonstrate the validity of this approach based on a combination of weights of attention extracted from the deep learning model and the linguistic knowledge stored in the WMAL lexicon. These considerations allow us to regard the approach provided in this paper as a promising starting point for further works in this research area.

1. Introduction

Machine learning and deep learning techniques have become increasingly popular in research on natural language processing topics. It is indeed easy to observe how almost all the tasks related to text classification, like tag annotation or question answering, have been recently addressed through approaches based on artificial intelligence. Although these new models have proven to be more effective and robust than their counterparts based on lexical analysis, they unfortunately have several limitations with respect to explainability. In many application areas, such as those related to eHealth, Economics, or Advertising, explainable models are preferred to black-box ones, even the fact that they usually offer worse performances and lower scores in terms of accuracy (Burkart and Huber, 2021). In those domains, trust is indeed the fundamental trait that makes the output of the models effective, valuable and relevant for the final user. For instance, in the eHealth scenario, no patient will trust an automatic diagnosis system that is not able to explain in detail the contribution of each symptom considered to build the diagnosis (London, 2019). Similarly, no user would confidently follow an automatically generated purchase suggestion without knowing whether it is based on purely commercial purposes or specific preferences extracted from their interests (Zhang, Chen et al., 2020).
Approaches for dealing with the explanation problem, such as LIME (Ribeiro, Singh and Guestrin, 2016) and SHAP (Lundberg and Lee, 2017), have recently been proposed in the literature. They are based both on the idea of training a

marco.polignano@uniba.it (M. Polignano); valerio.basile@unito.it (V. Basile); pierpaolo.basile@uniba.it (P. Basile); giuliano.gabrieli@crea.gov.it (G. Gabrieli); marco.vassallo@crea.gov.it (M. Vassallo); cristina.bosco@unito.it (C. Bosco)
ORCID(s): 0000-0002-3939-0136 (M. Polignano); 0000-0001-8110-6832 (V. Basile); 0000-0002-0545-1105 (P. Basile); 0000-0002-8857-4484 (C. Bosco)

second black-box neural model for the purpose of explaining the output of the first one. Nevertheless, such an articulated process can only be unsatisfactory since there will always be the need to explain the output of the secondary model. Moreover, the explanations are not generated by these approaches on the basis of the algorithm, but rather they try to estimate the behavior based on hypothetical descriptive features. However, situations where different explanations may be associated with the same example can occur simply because the model of explanation is invoked at different time instants (Angelov and Soares, 2020). Therefore, more recent research efforts (Ras, van Gerven and Haselager, 2018) highlight the need to use more transparent models, or forms of explanation that directly derive from the original model, rather than being created ad-hoc *a posteriori*. The application of this strategy allows for the design of more transparent and explainable models that are easier to understand to both the developer and the end-user.

Natural Language Processing approaches are historically been explained through the use of text analysis and explicit lexical, syntactic, and semantic features. Nowadays, approaches based on transformer architectures (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017) have become the state of the art. Those approaches are not transparent at all and their results difficult to explain. Indeed, transformer models are inherently black-boxes, and it is a very challenging task to explain their internal parameters to the final user.

Several attempts to turn the transformer architecture into an explainable model can be found in the literature. As an example, van Aken, Winter, Löser and Gers (2019) propose a detailed analysis of the individual levels of a transformer model, specifically for the question answering task. Unfortunately, what they obtain is a purely mathematical analysis that, no matter how useful to explain the operations of the algorithm, it does not provide an understandable explanation of the results achieved with the algorithm for the end-user. Vig (2019) proposes a tool to graphically visualize how the attention levels of a transformer model works during prediction. Although the tool introduces excellent graphical support for estimating the importance of each word with respect to the final classification label, this explanation is too complex and challenging to be understood by the end user. A similar direction is also taken by Bacco, Cimino, Dell'Orletta and Merone (2021b), who apply a summarization technique to the contextual representations generated by the transformer model in order to provide to end users a motivation for the polarity detection task results.

As a consequence of the growing demand and interest in the literature for higher performing and more explainable models but also explainable models in natural language processing, we decided to propose our hybrid BERT-WMAL model that merges lexical information with contextual representations generated by the transformer model. In doing so, not only do we obtain a model that is as performing as those currently at the state-of-the-art, but it also allows us to provide an explanation to the end user by exploiting the information we encoded through the WMAL component. In particular, we decided to define a two-way hybrid deep neural model. On the one hand, the textual input is transformed into a vector format (i.e., contextualized word embedding) through a state-of-the-art transformer model such as BERT. Conversely, the same input is transformed into a vector format by encoding the polarized words using the WMAL lexicon. These representations are then used in a deep neural network to perform the task of polarity detection. The presence of information from a polarized lexicon in the model not only supports the network in its pre-posed task but also provides a basis for the creation of a *a posteriori* explanation strategy. In particular, by appropriately varying the values associated with each term via WMAL, it is possible to estimate how important each of the term considered is relevant for the prediction function used by the network.
In a nutshell:

1. we propose the WMAL approach as a technique for building a weighted affective lexicon;
2. we provide a hybrid model which merges a transformer-based architecture with the WMAL lexicon;
3. we discuss a strategy for deriving an explanation of the model;
4. we conduct an exhaustive experiment and an end-user user study for supporting the validity of the proposed approach.

The rest of the article is organized as follows: Section 2 describes some related work in the literature; Section 3 presents the adopted methodology for WMAL; Section 4 discusses the hybrid BERT-WMAL model and its experimental results; Section 5 presents the explanation strategies and the findings emerging from the user study; Section 6 describes the limits and the challenges of the work, while conclusions and future work are presented in Section 7.

1.1. Research Goals

The main research aim of this paper is to investigate the possibility of creating a polarity detection model that can take advantage of the latest machine learning models but at the same time exploits the knowledge explicitly expressed in a polarized lexicon for SA. In particular, the experimental hypothesis that we believe can be addressed by this work

is that the inclusion of external information in a deep learning model can not only be helpful to improve the model performance, but also be used to design an explanation process that is effective for human users. Moreover, given the importance of having effective models also for languages less widespread than English, we decided to face this challenge with models developed for the Italian language.

Therefore, this work aims to address the following research challenges:

- Investigate the feasibility of using a model based on transformer architectures for the polarity detection task in the Italian language;
- Evaluate the possibility of using a polarized lexicon as a source of information external to the model to improve its effectiveness;
- Define a process for building a hybrid polarity detection model that can merge transformer architecture-based models and polarized lexicons;
- Investigate the possibility of making the designed hybrid model explainable;
- Evaluate model performance and different explanation strategies to identify the one considered most effective by humans.

The novelty of the work is supported by the great interest of the research community working on Artificial Intelligence (AI) in including external knowledge for supporting the explainability of models. In this work, not only do we propose a novel affective lexicon, but we also seek to demonstrate that its use in conjunction with the newest Transformers deep learning models can support both performances and explanation. Differently from other works in the literature, we therefore propose a hybrid model that supports data and external knowledge for improving the classification performances and uses the same information for providing a post-hoc explanation, applying a strategy based on the concepts of "contextual importance" (CI) and "contextual utility" (CU).

2. Related Work

Sentiment Analysis (SA), often mentioned in the literature as Opinion Mining (OM), is the task that in computer science aims to extract and analyze people's opinions, sentiments, attitudes, perceptions, etc., toward different entities considered as targets, such as topics, products, and services (Birjali, Kasri and Beni-Hssane, 2021). It has been a topic of research interest in Natural Language Processing for a number of years (Liu, 2012). SA can be performed with different levels of granularity as it is possible to detect the sentiment values and polarities about the whole document, single sentences, or specific aspects mentioned in them. The first two levels of granularity are among the mostly faced in literature. Still, the last one is currently of great importance due to its direct consequences on the task of comment and review analysis widespread in different industrial settings, e.g., to drive financial strategy.
In this work, we decided to approach SA over data at the sentence level and, in particular, by addressing the task of polarity detection over contents generated by users of Social Media platforms (Dridi and Reforgiato Recupero, 2019). In this article, we adopt the term "polarity detection", which better specifies the exact nature of the SA tasks we carry out, i.e., the classification of the polarity of the sentiment expresses by a natural language expression on an ideal scale from negative to positive. This term allows us also to differentiate our approach with respect to other instances of SA, such as Aspect-Based SA (ABSA) or emotion analysis and detection.

In the literature, the polarity detection challenge is addressed from many points of view, e.g. classifying polarity according to different granularity levels and scales of values, and applying approaches based on different knowledge sources and/or principles and strategies.

As regards **granularity and values used for polarity classification**, in the work proposed by Okanohara and Tsujii (2005) the polarity is calculated as a continuous number in the range [-1,1] where -1 indicates the negative polarity, 0 the neutral, and 1 the positive one. In other works, it is instead divided into classes (Qazi, Raj, Hardaker and Standing, 2017). As an example, in (Sharma, Nigam and Jain, 2014), three classes are used, i.e., positive, negative, neutral. El-Din (2016) instead proposed to use five different classes: very negative, negative, neutral, positive, very positive. In this work, we follow the polarity schema used in the dataset that we exploit for the evaluation of the proposed approach, i.e., the one based on four polarity classes: positive, negative, neutral, mixed (Barbieri, Basile, Croce, Nissim, Novielli and Patti, 2016).

Concerning **approaches used for detecting polarity**, a sharp classification of three categories can be found in the literature: Lexicon-Based approaches, Machine Learning approaches (supervised and deep learning), and Hybrid approaches (Birjali et al., 2021).

The **lexicon-based approach** (also called knowledge-based) is one of the most widely used strategies. This category of solutions is among the earlier adopted because it is straightforward and inherently explainable. A lexicon-based SA model exploits a lexical resource called polarized lexicon, i.e., a list of words annotated against a score of negative or positive polarities (Jurek, Mulvenna and Bi, 2015; Lahase, Shelke, Jagdale and Deshmukh, 2022). The score reflects the strength and the intensity of the sentiment expressed by the word. The final polarity of a document or sentence is obtained by summing the polarity values of every word composing the text. Some pre-processing steps must be applied to the text to allow a careful extraction of the words that occur in the document. These words are obtained through a process commonly known as tokenization. It aims to identify the component words of a text by dividing it according to blank space character and some more complex decision rules, often learned through machine learning strategies (Kudo and Richardson, 2018). Lexicon-based approaches are effective in well-constrained application scenarios and domains. This issue is a consequence of the difficulty of producing by manual annotation huge polarized lexicons. Indeed, they are often limited in size, and they contain only the words of the application domain. General-purpose lexicons generated with automated procedures are feasible but less frequent due to their low accuracy.

While lexicon-based methods are employed in specific settings, the vast majority of approaches proposed in the scientific literature are currently **approaches based on supervised learning and deep learning** in particular (Agüero-Torales, Salas and López-Herrera, 2021; Rosenthal, Farra and Nakov, 2017). A Machine learning approach can learn and generalize patterns from text, leading to best classification results also in domains other than those ones on which the approach has been trained. Nevertheless, in this kind of approach, the achievement of good performances requires training on very large dataset. Applying deep learning models to SA is a very common process. Deep learning, originated as a branch of machine learning, can offer models for supervised or unsupervised approaches. Their characteristic is the possibility of being able to generate representations of the input content that are increasingly abstract as the size of the network grows. This allows the network to identify complex relationships among features that cannot be expressed in the linear or polynomial form of low rank. Such networks point back to human neural architectures where a similar approach to information management is applied (Seema et al., 2022). Such approaches thus allow for training more complex models on a much larger dataset and thus, produce state-of-the-art results in many application domains, including several tasks of natural language processes. Among them, Transformer models Vaswani et al. (2017) are currently those achieving the best performance, for example, according to the General Language Understanding Evaluation (GLUE) benchmark for tasks in the English language[1]. For the Italian language, the scenario is similar, with the state of the art on polarity detection being achieved by neural language models (Polignano, Basile, Basile, de Gemmis and Semeraro, 2019c; Pota, Ventura, Fujita and Esposito, 2021).

Hybrid approaches have been proposed for several tasks, including SA. They aim to combine information from a lexicon and that obtained through machine learning, thus associating the advantage in performance given by the power of lexical analysis with the generality provided by machine learning approaches. This allows, in particular, to cope with the ambiguities of terms by integrating context by using emotional terms [252]. Therefore, even for complex tasks, such as polarity detection, stability and effectiveness can be ensured (Gupta and Joshi, 2020). Several approaches in this line employ lexicons to extract additional features from natural language, and concatenate them to other engineered features (Deshmane and Friedrichs, 2017). As an example, Li, Zhu, Shi, Guo and Cambria (2020) take into consideration the sentiment lexicon information to pad the sentence and to make the input data sample of a consistent size and to improve the proportion of sentiment information in each review. Then CNN-LSTM and CNN-BiLSTM models were proposed for obtaining the final sentiment score of the review. Extensive experiments demonstrated that sentiment lexicon information and parallel two-channel models could meaningfully improve the accuracy of SA models. With newer neural models computing lexical representations at training time, this approach is becoming less popular, but lexicons continue to be employed in larger supervised models for knowledge transfer across domain (Pamungkas and Patti, 2019), or to boost the performance of transformer models (Koufakou, Pamungkas, Basile and Patti, 2020). Motivated by the above discussion, in this paper, we attempt to propose a hybrid model for accomplishing a classification task that benefits from the exploitation of a polarized lexicon and also uses the knowledge available in this resource to support the explanation of the generated output and predictions about the sentiment expressed in textual data.

[1] https://gluebenchmark.com/leaderboard accessed on March 16th, 2022

The advantage of machine and deep learning methods consists in highly accurate predictions in conjunction with large dataset. Unfortunately, these methods work as black boxes and lack transparency and explainability by making it difficult to understand the internal features and representations of the data that a model uses to classify texts into specific sentiment categories.

In Artificial Intelligence, to address the natural opacity of approaches, there has been a development of many new methods to make models more explainable and their results more understandable for final human users: **Explainable AI** (XAI) (Samek, Wiegand and Müller, 2017; Samek and Müller, 2019; Vilone and Longo, 2020) aims especially to address how AI systems make decisions. The main idea is to define AI methods and techniques that produce human-comprehensible solutions for improving human understanding of the results generated by AI tools and models. XAI methods can indeed justify decisions made by the machine introducing trust and reducing bias. XAI is an emergent trend also in SA, where it can help to understand the features used by machines to classify texts according to different polarities. While lexicon-based and rule-based SA systems are explainable by themselves, existing XAI methods for supervised models can be applied to SA when it is performed as a classification task (Clos, Wiratunga and Massie, 2017). Several approaches are based on the idea of exploring attention-based techniques. An analysis of the performance of different attention-based approaches in SA is provided in (Bodria, Panisson, Perotti and Piaggesi, 2020). More sophisticated approaches try to combine Aspect-Based Sentiment Analysis (ABSA) with a classical Document Level Sentiment Analysis (DLSA). In (Silveira, Uszkoreit and Ai, 2019), the authors propose to average the polarity of each aspect derived by the ABSA model and use these aspects to explain the prediction performed by the DLSA model. The approach proposed in (Perikos, Kardakis and Hatzilygeroudis, 2021), e.g., by exploiting a Hidden Markov Model (HMM) into the training architecture, can highlight the portion of the sentence which especially contributes to determining the overall sentiment. Other approaches try instead to integrate external knowledge to improve the explainability of large neural language model results. Authors in (Zhao and Yu, 2021) propose a model based on BERT that can integrate external knowledge coming from a sentiment knowledge graph. This approach can improve interpretability and performance by incorporating domain-specific knowledge. In the context of SA of texts from reviews, the authors of (Bacco, Cimino, Dell'Orletta and Merone, 2021a) propose an approach able to extract summaries for explaining the predictions performed by a model based on Transformers. While the model proposed in (Arous, Dolamic, Yang, Bhardwaj, Cuccu and Cudré-Mauroux, 2021) relies on rationales provided by humans during the annotations of relevant pieces of text that explain the classification. In particular, human rationales are injected into a Bayesian framework that jointly learns an attention-based model and the reliability of annotations provided by humans. Recently, authors in (Cambria, Liu, Decherchi, Xing and Kwok, 2022) proposed a commonsense-based neuro-symbolic framework for building trustworthy symbolic representations able to extract polarity from a text in a completely interpretable and explainable manner. A more detailed overview of methods that rely on external knowledge for explainable machine learning, particularly knowledge graphs, is reported in (Tiddi and Schlobach, 2022). Finally, an extensive evaluation of machine learning approaches for SA of tweets is reported in (Fiok, Karwowski, Gutierrez and Wilamowski, 2021); the authors also evaluate several explainable artificial intelligence techniques to understand models' predictions.

3. Affective Lexicons for Italian

Affective lexicons are language resources where sentiment-related knowledge is encoded at the word-level. As such, affective lexicons are particularly suited for SA tasks such as the one presented in this study. A few affective available lexicons for Italian are known in the literature. The lexicon developed in the context of the OpeNER project (Maks, Izquierdo, Frontini, Agerri, Vossen and Azpeitia, 2014) contains 24,293 lexical entries manually labeled with positive/neutral/negative polarity. Sentix (Basile and Nissim, 2013a) contains, in its 2.0 version, 41,800 Italian lemmas with a polarity score ranging from -1 (totally negative) to 1 (totally positive), and derived from the automatic alignment of WordNet-based resources. EmoLex (or, the NRC Word-Emotion Association Lexicon) is a dictionary of emotionally-charged words with crowdsourced affective scores. Its Italian translation counterpart contains 9,921 entries organized around 8 basic emotions.

In the weighted version of an affective lexicon, the relative frequency of each used term is emphasized. Therefore, the weighted dictionary-based strategies are mainly able to classify the most sought-after words that substantially impact the interaction between sentiment and topic. Whenever people talk about a topic, particularly about a polarized topic, they presumably use even more sought-after words to better explain their points of views. The assumption that the less frequent terms should have a more substantial impact than the more frequent ones on polarity scores is the

basis for term weighting encoded in the affective lexicon WMAL, i.e., Weighted Morphologically-inflected Affective Lexicon (Vassallo, Gabrieli, Basile and Bosco, 2020).

3.1. The WMAL Lexicon

We initially proposed the Morphologically-inflected Affective Lexicon (MAL) (Vassallo, Gabrieli, Basile and Bosco, 2019). In this resource, each lexical item found within tweets with the entries of Sentix 2.0 is included as a lexical form, i.e. without applying any explicit lemmatization procedure. This lexicon was therefore expanded by including for each of the forms also all the other forms related to the same lemmas, as available in the Morph-It collection of Italian forms (Zacchetta and Baroni, 2006). As a result, each form gains the same polarity score of the original lemma, and when different lemmas may assume the same form, the arithmetic mean of their polarity scores was assigned. The MAL encompasses 148,867 forms and all the lexical items linked to the lemmas of Sentix 2.0. It can be therefore defined as an extension of Sentix itself, where the polarity scores are available for each inflected form related to each lemma originally occurring in Sentix.

Successively, we used TWITA, a large-scale corpus of messages from Twitter in the Italian language (Basile and Nissim, 2013b), as a reference weighting resource for extending MAL. TWITA is indeed both large (covering over 500 million tweets from 2012 to 2018, and the collection is currently ongoing) and domain-agnostic enough to yield a substantial representative sample of the distribution of the Italian language chatted words, although specific to Twitter only. However, and despite its size, not all the terms from the MAL occur in TWITA: only 57.9% of the 148,867 terms in MAL were found in TWITA due to the sparseness of particular inflected forms and to the presence of multi-word expressions in the lexicon (18,661, about 12%) that were not considered for matching the resources.

The scores of MAL were therefore recalculated by weighting them according to the word frequencies retrieved from TWITA and standardized using the Zipf scale measure (van Heuven, Mandera, Keuleers and Brysbaert, 2014). This measure was selected because of its easy understanding and the short computational timing. The Zipf scale measure is a logarithmic scale based on the well-known *Zipf's law* of word frequency distribution (Zipf, 1949). The computation of Zipf values of terms frequencies from TWITA essentially consisted of the logarithm of the frequency count per million words according to the original numerical expression[2] (van Heuven et al., 2014):

$$Zipf(i) = log_{10}\left(\frac{f(i)+1}{\frac{\sum_{i=1}^{N} f(i)}{10^6} + \frac{N}{10^6}}\right) + 3$$

where N is the number of tokens in the TWITA dataset (that is 6,644,867 tokens), $f(i)$ is the frequency count of the $i-$th token in TWITA, and the sum of the token frequencies $\sum_{i=1}^{N} f(i) = 6,906,070,053$; therefore:

$$Zipf(i) = log_{10}\left(\frac{f(i)+1}{6,906.07 + 6.644}\right) + 3$$

The original Zipf scale is a continuous scale ranging from 1 (very low frequency) to 6 (very high frequency) or even 7 (e.g., for especially frequent words like auxiliary verbs). The resulting weights in the weighted version of MAL, that is WMAL, ranged from a minimum of -5.16 to a maximum of 5.95 (the original MAL ranged from -1 to 1). For the lexical items that were not found in TWITA, in the WMAL we decided to keep the score originally associated in MAL.

Finally, to give proper weight to low-frequency terms, we straight reversed the Zipf scale by weighting the original scores inversely to their frequency. As reported in (Vassallo et al., 2020), we obtained promising results both in terms of improvement of the performance in prediction and stability across the positive and negative polarity classes, especially for more specific topics.

4. A Hybrid Model for Polarity Classification

The early research approaches concerning SA, and more precisely polarity classification, were based mainly on lexical resources, following the idea that are the words belonging to some particular grammatical category, such

[2]The original formula considers the two constants +1 and +3 to respectively take into account unobserved word types and words with frequencies of 1, or less, per million words with a Zipf value of 3 or less. However, by computing the Zipf score of the terms included in MAL according to their frequency in TWITA, we found some terms with a shallow frequency, thus yielding negative values of the logarithmic function. These were re-coded with the minimum Zipf value found in our calculation.

as verbs, adjectives, and adverbs, that mostly express the sentiment provided by a given sentence (Liu and Zhang, 2012). In this regard, several lexical resources were proposed such as SentiWordNet (Baccianella, Esuli and Sebastiani, 2010), NRC Affect Intensity Lexicon (Mohammad and Kiritchenko, 2018), and WordNet Affect (Strapparava, Valitutti et al., 2004). Following a straightforward approach, the cumulative polarity of a sentence is calculated as the simple sum of the polarity values associated with affective tokens occurring in it. Instead, the classifiers exploited by more complex approaches use more numerous and more informative linguistic features as inputs, including the TF-IDF (Term Frequency - Inverse Document Frequency) of the affective words occurring in the sentence, n-grams, punctuation, hashtags, and emoticons (Maas, Daly, Pham, Huang, Ng and Potts, 2011). Following this path, approaches based on Naive Bayes classification algorithms, Support Vector Machine (SVM), Decision Trees and Random Forest, Multilayer-perceptron, etc., have been proposed in the literature (Nayak and Natarajan, 2016). Although they have led to promising and encouraging results, they are not without limits yet to be addressed. Lexicons, in fact, have the disadvantage of not being easy to define, update and validate, and the context of use should be taken into account. For instance, the polarity of a word can vary according to its role and use in the sentence. Such lexicon-based approaches subsequently can provide excellent performance with lexical and syntactic features, but they do not do the same with the semantic ones.

With the aim of overcoming this semantic gap, new approaches based on word embeddings have emerged as state-of-the-art systems (Giatsoglou, Vozalis, Diamantaras, Vakali, Sarigiannidis and Chatzisavvas, 2017). In particular, a word embedding can be considered as a numerical, dense, vectorized representation of a single word or an n-gram, such that the semantics of the terms is preserved in the multidimensional space in which it can be projected. We obtain a distributional space in which interesting semantic properties emerge. In particular, observing terms as points in the semantic space, we can see that the more two terms are semantically similar and related, the closer they will be. Moreover, sum and addition operations between word embeddings also allow us to generate meaningful vectors from the semantic point of view. This technique will enable to obtain compact representations not only of single terms but also of whole sentences and documents by simply creating centroid vectors for them.

Word embeddings have been widely used in SA, often in conjunction with deep learning techniques, particularly with recurrent neural networks (RNN and LSMTM) (Polignano, Basile, de Gemmis and Semeraro, 2019b). Nevertheless, this representation technique is susceptible to the problem of contextuality. In a Word2Vec (Mikolov, Grave, Bojanowski, Puhrsch and Joulin, 2018) or GloVe distributional space (Pennington, Socher and Manning, 2014), each term is represented by a single embedding vector, whatever its context of use. In order to overcome this limitation, in the literature contextualized word representation techniques have emerged, generating word embeddings according to the sentence in which the words are used. Among these, we can mention Universal Sentence Encoder (Cer, Yang, Kong, Hua, Limtiaco, John, Constant, Guajardo-Cespedes, Yuan, Tar, Strope and Kurzweil, 2018), ELMO (Peters, Neumann, Iyyer, Gardner, Clark, Lee and Zettlemoyer, 2018), BERT (Devlin, Chang, Lee and Toutanova, 2019), RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer and Stoyanov, 2019), XLM (Conneau, Khandelwal, Goyal, Chaudhary, Wenzek, Guzmán, Grave, Ott, Zettlemoyer and Stoyanov, 2020). BERT is certainly the most common and famous strategy among them.

BERT (Bidirectional Encoder Representations from Transformers), in its basic version, is trained on a Transformer network with 12 encoding layers, 768 dimensional states and 12 heads of attention for a total of 110M of parameters trained on BooksCorpus and Wikipedia English for 1M of steps. The learning phase is performed by scanning the span of text in both directions, from left to right and from right to left, as was already done in BiLSTMs. Moreover, BERT uses a "masked language model": during the training, random terms are masked in order to be predicted by the net. Jointly, the network is also designed to potentially learn the next span of text from that one given in input. These characteristics allow BERT to be the current state-of-the-art language understanding model.

By following what has been recently proposed in the natural language processing state-of-the-art, we decided to propose a hybrid approach for detecting polarity. In particular, we supposed that we could combine the expressive power of a lexicon with the semantic properties of a contextual embedding representation based on BERT. In fact, we hypothesize that such an approach can not only increase the predictive capacity of the model, but also support with explanations the end-user, allowing them to understand the motivations beyond the specific polarity assignment (positive, negative, mixed, neutral class) to each given example the model does.

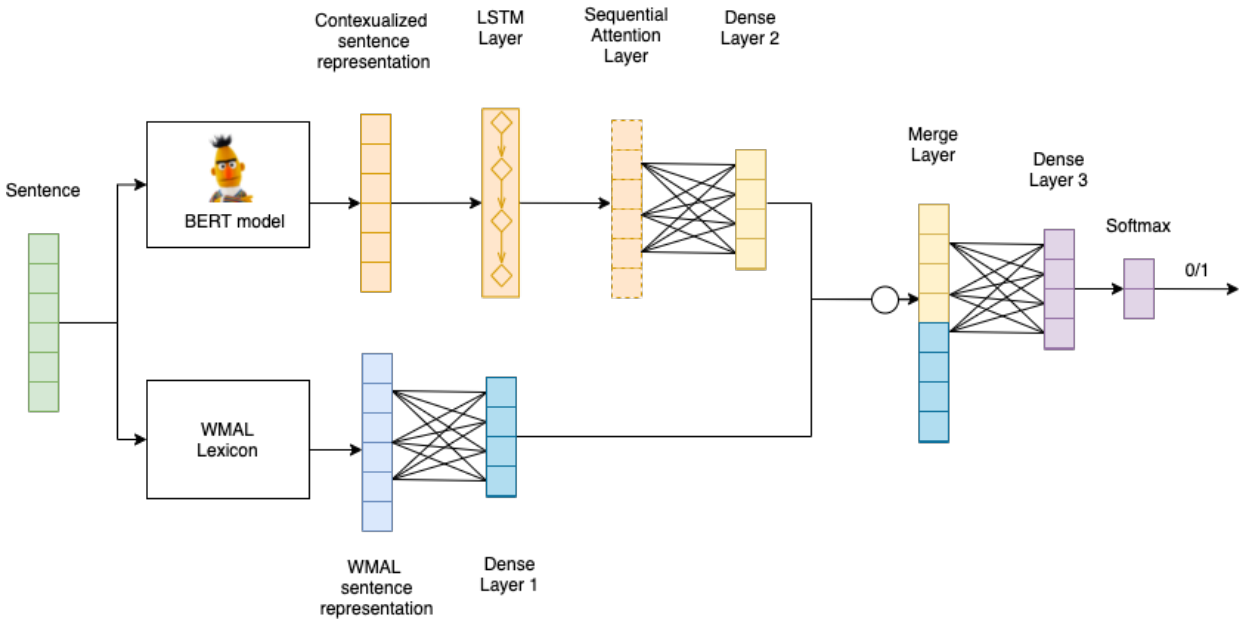


Figure 1: Hybrid BERT-WMAL model proposed. The input text is encoded in two parallel branches, respectively by BERT followed by an LSTM and a dense connected layer, and by WMAL followed by a dense connected layer. The outputs of the two branches are concatenated before the final prediction.

4.1. The model

The architectural model we propose combines two components in a single deep architecture, i.e., the one derived from the BERT based representation and the other given by the WMAL encoding. Specifically, the design of this model is shown in Fig. 1, where the main steps of its architecture are provided. In particular, the sentence to be classified is supplied as an input to the model after being opportunely pre-processed. We decided to use a classical pre-processing pipeline such that punctuation, non-alphanumeric elements, and repeated spaces were removed, and the remaining tokens were taken in their lowercase form. These pre-processed sentences were then provided in parallel as an input to both the BERT model and the WMAL-based text representation module.

Following the upper branch of the chart represented in Fig. 1, the BERT-based sentence representation initially performs tokenization of the text by applying the WordPiece strategy (Wu, Schuster, Chen, Le, Norouzi, Macherey, Krikun, Cao, Gao, Macherey, Klingner, Shah, Johnson, Liu, Kaiser, Gouws, Kato, Kudo, Kazawa, Stevens, Kurian, Patil, Wang, Young, Smith, Riesa, Rudnick, Vinyals, Corrado, Hughes and Dean, 2016). TFirst, this strategy tries to relate each word to some term occurring in the vocabulary of the BERT model. If this process is not successful, the algorithm will try to subdivide the word into the conjunction of its subparts. This process makes the tokenization process robust, fast, and versatile and allows to properly handle situations where a term provided as input does not occur in the model's vocabulary.

The tokens generated after this step are transformed into a vector form and suitably combined with a positional array that indicates their position within the sentence context. At this point, the input sentence is propagated in the network. Since the BERT model is based on a transformers architecture made by the succession of encoding modules, at the end of each module, the model provides an embedding representation for the tokens of the sentence. Different opinions have been proposed in the literature regarding how many encoding layers it is necessary to activate in order to obtain valid embeddings. According to (Devlin et al., 2019), a reasonable choice is to use the sum of the last four encoding layers as contextual word representations. Following this strategy, at the end of this process, we obtain a sentence representation of 128×768, i.e. the number of input tokens and the number of BERT hidden units.

As far as the lower branch of the architecture provided in Fig. 1, a straightforward process is used for WMAL-based text representation. In particular, the text has been tokenized by using the space as a separator. At this point, a vector characterized by the length that matches the number of tokens obtained was created. For each token, if it occurs in the WMAL lexicon, the corresponding score is used to fill the corresponding cell of the vector. Moreover, the cell is

initialized with the value 0. Over the WMAL-based representation, we add a Dense Neural Layer of 64 hidden units to reduce its size and make it more stable and comparable with other data available in the model. Over the BERT-based contextual representation, we decided to use a Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) layer in its bi-directional version. Considering the intrinsic sequential relationship between the terms of a sentence, i.e., the next term depends on the context composed of previous terms, the contribution made by a recurrent neural network in order to grasp this relationship is obvious. LSTM uses the forget gate (hidden neuron) to dynamically scale the weights of its internal "self-loop" depending on the weights learned by the network for previous words provided as input. This step provides the layer a "memory" for considering the relations with the past elements in the input.
We decided to apply an attention level to the weights estimated from the LSTM level. Specifically, in order to obtain a correct prediction, the attention function tries to learn which features are more significant than others. In this process, it is decided to assign different weights to different features according to their importance.

The attention layer and the WMAL representation are then concatenated through the *"Merge Layer"* and reduced in dimensionality across a dense layer, as shown in the right part of Fig. 1, where the two branches of the graph merge before the generation of the final prediction. Finally, a layer with SoftMax optimization function (Goodfellow, Bengio and Courville, 2016) is used to obtain the probability distribution with respect to two individual classes in the output. In particular, the model has been trained on the two polarities, positive and negative, independently. For each one, there will be a prediction in class 0, 1 indicating whether or not the model believes that the sentence is positive/negative. This will allow us to obtain 4 possible combinations of polarity: positive $(1, 0)$, negative $(0, 1)$, mixed $(1, 1)$ and neutral $(0, 0)$. We will discuss in depth every single configuration of the proposed model in the Chapter 4.2.

To sum up the key points about the process and the proposed model are:

- **Pre-processing**: removal of punctuation, non-alphanumeric elements, repeated space. Text transformed in lower case.
- **Generation of BERT embeddings**: Italian language BERT model, i.e. AlBERTo small version (L=12, H=748, A=12, uncased)[3]. Sum of last four layers. 128 tokens length. 128×768 shape.
- **Generation of WMAL embeddings**: 128 tokens length vectors with 0 if WMAL value not found for the corresponding token, the value otherwise.
- **Model**: Keras Functional API for a two-input model. Hyper-parameters of the different configurations evaluated are reported in the following.

The source code of the proposed model and related resources are publicly released through our GitHub repository[4].

4.2. Evaluation

We evaluate the model on three publicly available benchmarks for the Italian language. The first is the dataset released for the *SENTIPOLC* (SENTIment Polarity Classification) shared task carried out at EVALITA 2016 (Barbieri et al., 2016), a challenge on polarity detection on Italian tweets. The second is the AGRITREND (Vassallo et al., 2019) dataset which includes Italian tweets about the agriculture domain that were manually annotated with their polarity by three different annotators. The last is a portion of the *ABSITA* dataset (Basile, Croce, Basile and Polignano, 2018) released for a shared task of Aspect Based Sentiment Analysis at EVALITA 2018.

Since the datasets are in Italian, we opted for a version of BERT trained for the Italian language. In particular, we used AlBERTo (Polignano, Basile, De Gemmis, Semeraro and Basile, 2019a), a BERT-based model trained on tweets, not case-sensitive, with 12 layers, 768 hidden units, and 12 attention heads with 110M total parameters. We decided to use AlBERTo as the transformer model because it is currently the only BERT model for the Italian language generated and released by a research group and validated through a scientific publication on it. Moreover, it was developed specifically for the language of social media, and then it perfectly fits the data we considered in this work.

[3] m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0
[4] https://github.com/marcopoli/bert_wmal_model

Table 1
Descriptive statistics about the datasets used in the experimental session.

	# Train set	# Test set	Labels Pos	Labels Neg	Polarity
SENTIPOLC	7410	2000	0, 1	0, 1	positive, negative, neutral, mixed
AGRITREND	/	1000	0, 1	0, 1	positive, negative, neutral, mixed
ABSITA	2365	1171	0, 1	0, 1	positive, negative, neutral, mixed

Two deep neural networks were trained for each model configuration, one for each of the two distinct labels: positivity and negativity.

Datasets. The *SENTIPOLC* datasets, respectively provided for training and testing, are tagged with six fields containing values related to manual annotation applied on them: *subj*, *opos*, *oneg*, *iro*, *lpos*, *lneg*. These labels respectively indicate whether the sentence is subjective, positive, negative, ironic, literally positive, and literally negative. For each of these classes, the annotation is 1 where the sentence satisfies the label, 0 otherwise. The last two labels "lpos" and "lneg", which describe the literal polarity of the tweet, have not been considered in the current evaluation (nor in the official shared task evaluation). In total, 7,410 tweets have been released for training and 2,000 for testing.

Similarly *AGRITREND* is annotated with two labels *opos* and *oneg*, which indicate whether a tweet is positive or negative. It is important to note that, in case opos and oneg are both 0, the tweet is considered as neutral, while when both the categories are annotated as 1, a mixed polarity features the tweet. In total *AGRITREND* provides annotations for 1,000 tweets, and all of them are used as a test set.

ABSITA releases a polarity score (i.e., *pos* and *neg*) for each of the eight different aspects studied in the context of the challenge where the data were released: cleanliness, comfort, amenities, staff, value, wifi, location, and other. Data have been collected from the well-known website *Booking.com* and, in particular, including reviews about a random selection of hotels located in some Italian cities (i.e., Naples, Bologna, Milan) left by users in Italian. In this work, we focused on the "comfort" aspect because it is the most discussed by users, with 2,365 occurrences for training and 1,171 occurrences for testing. For our purposes, we are considering the whole review without focusing on the span of text containing the specific aspect.

The statistics about the three different datasets used in the experimental session are reported in Table 1.

Experimental protocol. We decided to evaluate different versions of our model by varying few layers at each time. We trained two models for each configuration (i.e. one for *pos* and one for *neg*). They were implemented by using Keras and PyTorch and were trained for 8 epochs with a batch size of 128, Adam optimizer with a learning rate equal to $5e{-}5$, and a categorical cross-entropy as a loss function. The number of epochs has been set after different preliminary attempts in a range between 3 and 10 by observing the loss score obtained. We decided to stop the model at a step where the loss function was around 0.04 to obtain a prediction accurate enough to avoid overfitting. The training phase was performed on the Google Colab platform using their TPU. Among the corpora cited above, in particular, we considered as training dataset that one developed for the SENTIPOLC task, while we run the test phase on three datasets, i.e., SENTIPOLC test set, AGRITREND, and ABSITA test sets.

Config 1. The specific configuration we decided to implement to be used in this task is the *"Sequential Attention Layer"* as a *Sequential Weight* type attention function proposed in (Felbo, Mislove, Søgaard, Rahwan and Lehmann, 2017). It's formal representation is reported in Eq. 1-3.

$$e_t = h_t w_a \tag{1}$$

$$a_t = \frac{exp(e_t)}{\sum_{i=1}^{T} exp(e_i)} \tag{2}$$

$$v = \sum_{i=1}^{T} a_i h_i \tag{3}$$

where h_t is the input representation at the step t and w_a is the weight matrix of the attention layer. The attention importance scores for each step, a_t, are obtained by multiplying the representations with the weight matrix and then normalizing the scores by constructing a probability distribution over the representations. Lastly, the representation vector for the input sentence, v, is found by a weighted sum over all the steps by using the attention importance scores as weights.
The *"Dense Layer 2"* is removed and then we exchange the *"LSTM Layer"* with its bi-directional variant. It considers the relations among inputs in both the directions, finally it provided as output the concatenation of the relations from both the sides as in Eq. 4.

$$x_i = \overleftarrow{x_i}||\overrightarrow{x_i} \quad x_i \in R^{2d} \tag{4}$$

where $||$ is the operator of concatenation and d is the dimension of the LSTM in terms of hidden units. We have configured the LSTM network by setting to 32 the value of hidden units. This choice was motivated by the need to reduce the dimensionality of the output of the network so that the operations carried out by the following layers were not computationally too expensive.

Config 2. In this model configuration, we decided to use as *"Sequential Attention Layer"* a simple multiplicative sequential attention function (Zheng, Mukherjee, Dong and Li, 2018) able to capture the similarity of any token with respect to all the neighboring tokens in a given input sequence. We reported its formula in Eq.5.

$$e_{t,t'} = \sigma(h_t^T W_a h_{t'} + b_a) \tag{5}$$

where σ is the element-wise sigmoid function, h_t is the token representation at the step t, h'_t is the token representation at the step t', w_a is the weight matrix of the attention layer and b_a is the optional bias score. The *"LSTM Layer"* is configured with 64 hidden units and the *"Dense Layer 2"* is removed.

Config 3. Configuration three is very close to configuration two. Also in this case, a multiplicative sequential attention function has been used, and the *"LSTM Layer"* is configured with 64 hidden units. The most significant change is about the *"Dense Layer 2"* that has been substituted with a *"Multiplicative Layer"*. In particular, the output of the *"Sequential Attention Layer"* has been multiplied with the output of the *"LSTM Layer"* in order to try to improve the quality of prediction.

Config 4. The fourth configuration is a slight variation of the third. In particular, we decided to use an *"Additive Layer"* between the output of the *"Sequential Attention Layer"* and the *"LSTM Layer"*. Then the *"Dense Layer 2"* has been kept for reducing the output the *"Sequential Attention Layer"* to 64 hidden-units.

Config 5. The last configuration is equal to configuration four instead of the *"Dense Layer 1"* which has been used for expanding the WMAL vectorial representation to higher dimensionality. In particular, it has been set at $1,000$.

AlBERTo. This model configuration directly exploits the BERT-based representations in order to make predictions.

no WMAL. We decided to use the *"Config 4"* of our model without the use of the WMAL component.

Metrics and baselines. As evaluation metrics, we decided to use the classical precision, recall, and F1 measures in their macro average version (Raschka, 2014). We compared our best model configuration with the best-scored systems of the SENTIPOLC 2016 task of polarity detection (Barbieri et al., 2016) and with the systems proposed in (Tamburini, 2020) and in (Pota et al., 2021). The first is based on the use of GilBERTo[5], another BERT-based model for the Italian language also properly fine-tuned for the polarity detection task. The second one proposes a novel pre-processing strategy of data instead to reduce noise and errors while using a BERT-based language model, pre-trained on plain text and fine-tuned on pre-processed tweets of *SENTIPOLC*.

[5]https://github.com/idb-ita/GilBERTo

Table 2
Precision scores for the different configuration of the proposed model.

	AGRITREND			*SENTIPOLC*			*ABSITA*		
Model	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**
Config 1	0.77016	0.61180	0.69098	0.73492	**0.77597**	0.75545	0.92214	0.91191	0.91702
Config 2	0.74409	0.58569	0.66489	0.73285	0.68589	0.70937	0.92164	0.89844	0.91004
Config 3	**0.78288**	**0.62986**	**0.70637**	0.74649	0.74685	0.74667	0.91798	0.91554	0.91676
Config 4	0.74818	0.60381	0.67600	0.75675	0.71539	0.73607	**0.92476**	**0.91990**	**0.92233**
Config 5	0.67530	0.54076	0.60803	0.70172	0.68789	0.69481	0.90537	0.90680	0.90608
AlBERTo	0.64311	0.57640	0.60976	**0.77625**	0.75400	**0.76513**	0.91633	0.91205	0.91419
no WMAL	0.73307	0.60296	0.66802	0.74568	0.71223	0.72896	0.90309	0.91047	0.90678

Table 3
Recall scores for the different configuration of the proposed model.

	AGRITREND			*SENTIPOLC*			*ABSITA*		
Model	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**
Config 1	0.58253	0.53126	0.55690	0.66998	0.68778	0.67888	0.90635	0.91282	0.90958
Config 2	0.66296	0.55501	0.60899	0.72898	0.72457	0.72678	0.92189	0.90104	0.91146
Config 3	0.58394	0.54763	0.56579	0.72107	0.72760	0.72434	0.91452	0.91395	0.91423
Config 4	0.72154	**0.61634**	0.66894	**0.73358**	**0.74553**	**0.73956**	**0.92600**	**0.91485**	**0.92042**
Config 5	0.67173	0.54694	0.60934	0.69678	0.68493	0.69086	0.89540	0.88901	0.89220
AlBERTo	**0.77162**	0.61169	**0.69165**	0.71935	0.68075	0.70005	0.91913	0.91047	0.91480
no WMAL	0.72114	0.61570	0.66842	0.73574	0.72156	0.72865	0.90134	0.90352	0.90243

Table 4
F1 scores for the different configuration of the proposed model.

	AGRITREND			*SENTIPOLC*			*ABSITA*		
Model	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**
Config 1	0.61039	0.33591	0.47315	0.67460	0.71727	0.69593	0.91235	0.91235	0.91235
Config 2	0.65603	0.38390	0.51996	0.73070	0.70062	0.71566	0.92176	0.89956	0.91066
Config 3	0.57788	0.35710	0.46749	0.72803	0.73645	0.73224	0.91612	0.91470	0.91541
Config 4	**0.73219**	**0.60653**	**0.66936**	**0.74048**	**0.72820**	**0.73434**	**0.92536**	**0.91700**	**0.92118**
Config 5	0.67343	0.52388	0.59865	0.69600	0.68639	0.69119	0.89949	0.89466	0.89707
AlBERTo	0.70125	0.59352	0.64752	0.72912	0.71553	0.72232	0.91763	0.91122	0.91442
no WMAL	0.72649	0.60197	0.66423	0.73957	0.71667	0.72812	0.90218	0.90631	0.90424

4.3. Discussion of results

As the experimental results reported in Tables 2 - 5 show, the proposed model performs very well on the datasets used for evaluation. In particular, the observation of the F1 scores shows that "*Config 4*" clearly outperforms the results achieved using the other configurations, including the use of AlBERTo in its basic version trained with fine-tuning for the polarity detection task.

Focusing on the details reported in Table 4, it can be observed how both the insertion of an appropriate level of attention and the addition of information coming from the WMAL representation bring a positive contribution to the model's performance based according to the F1 score. By comparing *"Config 4"* with *"AlBERTo"*, it is possible to observe an increase in F1 score of 3.373%, 1.664%, 0.739% respectively, on the *AGRITREND*, *SENTIPOLC* and *ABSITA* datasets. While comparing *"Config 4"* with *"no WMAL"* we observe an increase in F1 score of 0.772%, 0.854%, 1.873% respectively, on the *AGRITREND*, *SENTIPOLC* and *ABSITA* datasets. Although WMAL provides a limited improvement of the score achieved in the classification task, it plays an essential role in the explanation phase of the obtained predictions. Therefore, we believe that the WMAL contribution not only succeeds in contributing to the model's performance but also enables its explainability.

Table 5
Accuracy scores for the different configuration of the proposed model.

	AGRITREND			*SENTIPOLC*			*ABSITA*		
Model	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**	**NEG**	**POS**	**AVG**
Config 1	0.75761	0.36917	0.56339	0.72550	**0.86100**	0.79325	0.91631	0.91375	0.91503
Config 2	0.78093	0.42698	0.60395	0.74700	0.80750	0.77725	0.92400	0.90094	0.91247
Config 3	0.75963	0.39148	0.57555	0.75400	0.85300	0.80350	0.91887	0.91631	0.91759
Config 4	**0.79513**	**0.63083**	**0.71298**	**0.76400**	0.83100	**0.79750**	**0.92741**	**0.91887**	**0.92314**
Config 5	0.73935	0.53854	0.63894	0.71800	0.81950	0.76875	0.90350	0.89838	0.90094
AlBERTo	0.77521	0.61756	0.69638	0.74750	0.82440	0.78595	0.91973	0.91289	0.91631
no WMAL	0.78499	0.62880	0.70689	0.75820	0.83200	0.79510	0.90521	0.90863	0.90692

Table 6
Comparison of F1 scores with the state of the art systems for SENTIPOLC 2016 dataset.

System	NEG	POS	AVG	AVG Δ
Config 4	**0.7282**	**0.7404**	**0.7343**	-
UniPI.2.c (Barbieri et al., 2016)	0.6426	0.6850	0.6638	-9.601%
Unitor.1.u (Barbieri et al., 2016)	0.6885	0.6354	0.6620	-9.846%
Unitor.2.u (Barbieri et al., 2016)	0.6838	0.6312	0.6575	-10.460%
ItaliaNLP.1.c (Barbieri et al., 2016)	0.6743	0.6265	0.6504	-11.430%
BERT Multilang (Polignano et al., 2019c)	0.4978	0.5511	0.5230	-28.780%
Pota et al. (2021) (Pota et al., 2021)	0.7671	0.7340	0.7506	+2.220%
Tamburini (2020) (Tamburini, 2020)	-	-	0.7475	+1.798%

By comparing the different configurations of the model, we can also observe how the introduction of a sequential multiplicative model of attention (*"Config 2"*) has allowed us to obtain much more satisfactory results than its weighted counterpart (*"Config 1"*). The idea of combining the output of the *"Sequential Attention Layer"* and the *"LSTM Layer"* instead rewarded the additive strategy (*"Config 4"*) rather than the multiplicative one (*"Config 3"*). Finally, the feature expansion strategy on the WMAL component did not lead to apparent improvements in the model performance. Taking into consideration precision and recall, it can be seen that, as far as *AGRITREND* is concerned, *"Config 3"* is the one with highest precision scores but lowest recall scores were achieved. For *SENTIPOLC*, a similar situation can be observed with the *AlBERTo* configuration. On the contrary, focusing on recall, it is possible to observe that "*Config 4*" behaves well on average on both datasets, which makes it the most stable configuration in terms of F1 score. About the *ABSITA* dataset "*Config 4*" obtains the best results for both precision, recall, F1 and accuracy scores.

By considering instead the baselines reported in Table 6, we can observe how the proposed model obtains results in line with state-of-the-art. In particular, it gets a better F1 score than the winning and best scored models of the SENTIPOLC 2016 challenge, with a difference in performance ranging from 9.601% to 11.430%. If we compare the proposed approach with the base version of BERT multilingual, we can instead observe a variation in terms of F1 score of more than 28%. A different result can be observed by comparing the best configuration of our model with the strategy proposed by Pota et al. (2021) and Tamburini (2020). In fact, these two models obtain a value of F1 score from 1.798% to 2.220% higher than those ones we observed. The main differences with the approach we proposed are about the pre-processing strategy and the BERT-based architecture chosen. In our work, we cleaned the texts from any non-alphanumeric characters, including punctuation and emoji. In literature about polarity detection, these elements are commonly known for their contribution to the correct classification of textual elements. Indeed, in the work proposed by Pota et al. (2021), a pre-processing strategy is applied which is based on the Ekphrasis library[6] and preserves emoji, punctuations, and many other linguistic elements. Ekphrasis is a collection of lightweight text tools geared toward text from social networks for tokenization, word normalization, word segmentation (for splitting hashtags), and spell correction, using word statistics from two big corpora (English Wikipedia, English tweets). The tokenizer can understand complex emoticons, emojis, and other unstructured expressions like dates, times, and more. The library is available for the Python programming language, and it results very effective when applied to textual contents extracted

[6]https://github.com/cbaziotis/ekphrasis

Table 7
Results obtained by the proposed model, i.e. Config 4, by varying the percentage of SENTIPOLC dataset splitting. The last two columns reported the average values (AVG) and the standard deviation among results (σ).

		60-40	65-35	70-30	75-25	80-20	85-15	90-10	AVG	σ
Precision	**POS**	0,65662	0,68074	0,69755	0,70251	0,71539	0,73302	0,73327	0,70273	0,02783
	NEG	0,73306	0,72933	0,73330	0,73327	0,75675	0,75000	0,75183	0,74108	0,01129
Recall	**POS**	0,62255	0,68892	0,67295	0,73260	0,74553	0,73186	0,71315	0,70108	0,04314
	NEG	0,72003	0,71173	0,73682	0,72924	0,73358	0,76288	0,75878	0,73615	0,01888
F1	**POS**	0,63234	0,68446	0,68258	0,71449	0,72820	0,73243	0,72191	0,69949	0,03573
	NEG	0,72429	0,71665	0,73480	0,73097	0,74048	0,75132	0,75456	0,73615	0,01377
Accuracy	**POS**	0,74681	0,75824	0,77400	0,79898	0,83100	0,78966	0,80553	0,78632	0,02898
	NEG	0,74283	0,73770	0,74531	0,74416	0,76400	0,75708	0,76621	0,75104	0,01126

from Twitter or, more in general, social media platforms Baziotis, Pelekis and Doulkeridis (2017). Differently, in work proposed by Tamburini, an Italian pre-trained model of the Facebook RoBERTa architecture has been used. As already known in the literature, this architecture is more efficient and accurate than the classic BERT model we used (Wang, Singh, Michael, Hill, Levy and Bowman, 2018). In the future, we reserve the possibility to investigate the influence of the pre-processing phase on the performance of the model proposed in this paper. Moreover, we will explore the use of different Transformer-based models.

4.4. Robustness Analysis

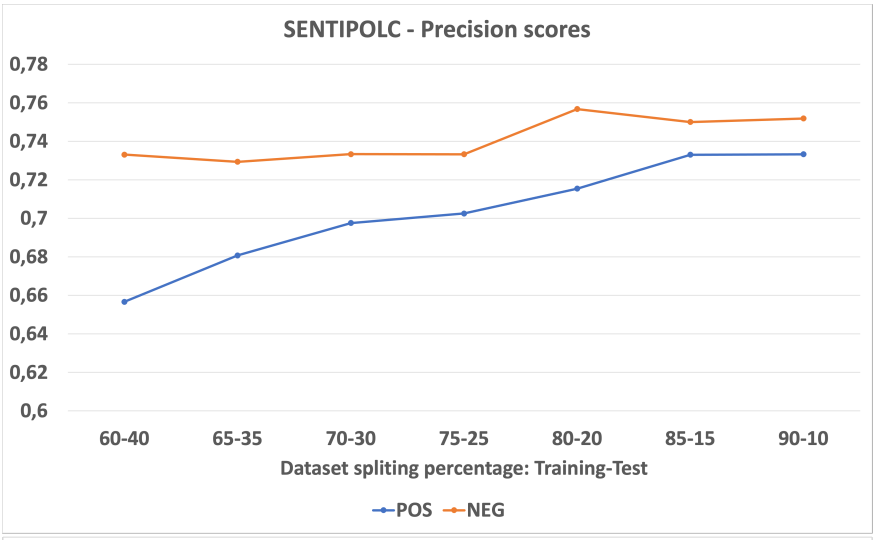
To verify the robustness of the proposed model, particularly that of the most performing configuration *(Config 4)*, we evaluated its behavior by varying the proportions of the training and test data. Specifically, the SENTIPOLC dataset in its entirety was taken as a reference, i.e., by merging the two training and test partitions initially provided and evaluating a percentage split of: 60-40, 65-35, 70-30, 75-35, 80-20, 85-15, 90-10.

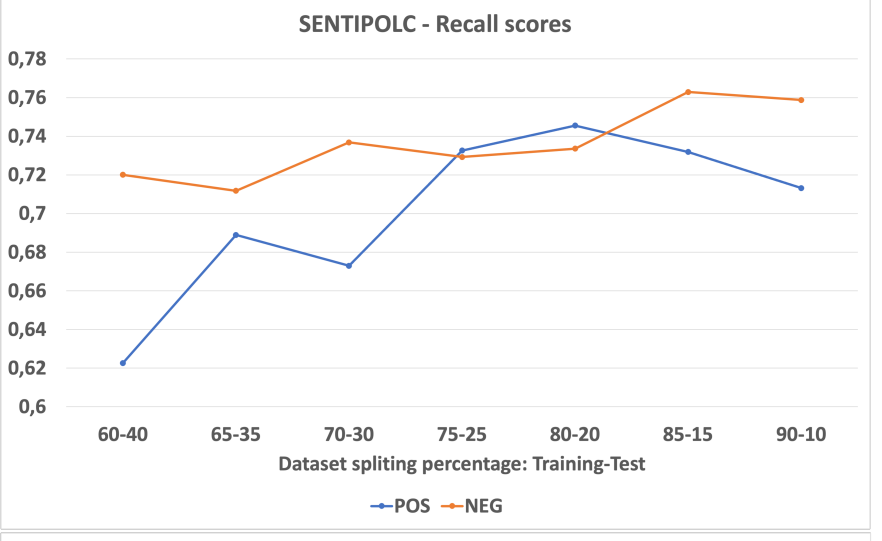
Table 7 and Fig. 3 resume our findings. In particular, we can observe that the model seems robust to the reduction of data provided for training. Indeed, the scores obtained by all the four different metrics (i.e., precision, recall, F1, and accuracy) are all confined to a small range of variation, which can be naturally considered as a consequence of the reduction of data available to the model for training a data representation general enough for the specific classification task. In support of this assertion, we can analyze the standard deviation value (σ), which takes on minimal values between 0.01126 and 0.04314. This denotes low variability in the results obtained and thus a good level of model robustness.

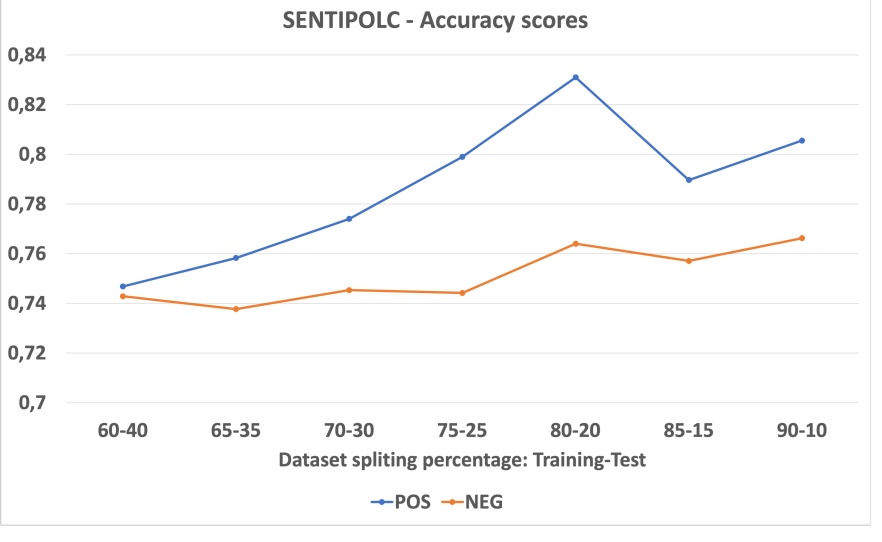
The graphs in Fig. 3 show higher robustness in the negative content class than in the positive content class. Indeed, the values of the different metrics obtained for that class are less sensitive to variation in training data quantities. Therefore, we can hypothesize that the model can learn how to correctly distinguish contents with negative polarity from the rest of the contents. This is probably also due to the strong influence on negative elements by WMAL. Indeed, it has already been shown in (Vassallo et al., 2019) that WMAL turns out to be particularly useful in connoting negative versus positive aspects. As further evidence, we can observe that the trend lines of Accuracy values, compared to F1 values for the POS class, are far apart. This indicates that the positive class is more unbalanced in the training phase of the model. This causes it to fail to generalize correctly and tends to often prefer the positivity of the class over its neutrality or negativity as a prediction. Nevertheless, the model achieves a reasonably high F1 score and is in line with the state-of-the-art scores achieved on the same dataset.

5. Lexicon-driven Classification Explanation

In Section 4, we have already observed how the hybrid model integrating WMAL with a BERT-based model obtains F1 scores comparable with those achieved for the same task by the state-of-the-art models. This confirms that the WMAL polarized lexicon can be usefully integrated with the embeddings of a transformer model and that this strategy does not negatively affect the classification ability of the learned model. Nevertheless, the real added value of the integration in our model of such polarity features mainly consists of improving the results' explainability. In our hybrid model, the actual contribution of each value of WMAL can be observed, and the relevance in the classification process of each term occurring in the input sentence is carefully evaluated. To reach this goal, we were inspired by the


SENTIPOLC - Precision scores
0,78
0,76
0,74
0,72
0,7
0,68
0,66
0,64
0,62
0,6
60-40
65-35
70-30
75-25
80-20
85-15
90-10
Dataset spliting percentage: Training-Test
POS
NEG


SENTIPOLC - Recall scores
0,78
0,76
0,74
0,72
0,7
0,68
0,66
0,64
0,62
0,6
60-40
65-35
70-30
75-25
80-20
85-15
90-10
Dataset spliting percentage: Training-Test
POS
NEG


SENTIPOLC - Accuracy scores
0,84
0,82
0,8
0,78
0,76
0,74
0,72
0,7
60-40
65-35
70-30
75-25
80-20
85-15
90-10
Dataset spliting percentage: Training-Test
POS
NEG

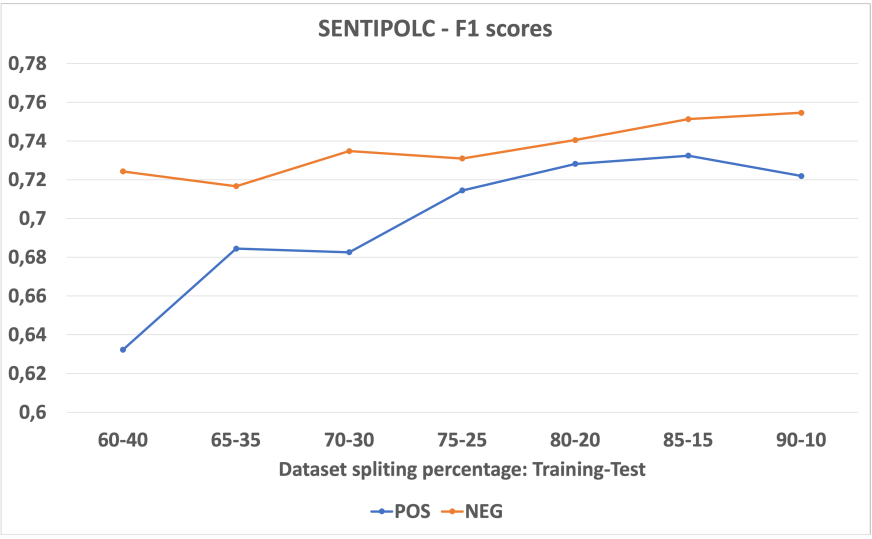


Figure 3: Graphical visualization of the variation in scores by varying the percentage of SENTIPOLC dataset splitting. It is possible to observe that all the scores naturally decrease by reducing the amount of training data. Generally speaking, the model is relatively stable and the variation is confined to a small range of differences in the scores.

work by Fouladgar et al. (Fouladgar, Alirezaie and Främling, 2020) and by what is commonly performed in the world of Multi Criteria Decision Making (MCDM) (Främling, 2020) domain. In particular, what is attempted to understand is how much a feature is *important* for the classification obtained and how much its value is actually *useful*, given the class. The two concepts we have taken as reference are therefore those of *"contextual importance" (CI)* and *"contextual utility" (CU)* over the WMAL model features. On the one hand, CI allows us to answer questions like "How important for the prediction given by the model is the WMAL value assigned to the specific sentence token?". On the other hand, CU supports us in answering questions like "Compared to the possible variety of probability values obtainable for the WMAL value, how good is the one assigned? Is there a better WMAL value for obtaining the predicted class with a higher confidence score?".

In order to obtain the scores of CI and CU, at the end of the classification step, we vary, for each sentence s_i of the test set, the WMAL value of each j-esim term $s_{i,j}$, 100 times. The $s_{i,j}$, is the WMAL value assigned at the textual token in position j of the specific sentence s_i considered. After this process we obtain $sent\_length \times 100$ new sentence representations $s_{i,j,k}$ with $k \in [1-100]$. Where $sent\_length$ is the number of textual tokens composing the sentence $s_{i,j}$. The 100 WMAL-generated scores allow us to obtain statistically relevant results for the following analysis. The new WMAL values were obtained from a Standard Gaussian distribution with mean μ given by the average of all the WMAL values in the training set and standard deviation 3σ, so that we had a 99.73% probability of obtaining new values in the WMAL original values range. Given the new sentence representations, we ran over them the proposed hybrid BERT-WMAL model, and we obtained for each of them the class probabilities. With these results, we can calculate CI and CU for each sentence token $s_{i,j}$ as follows (Eq. 6 and Eq. 7):

$$CI(s_{i,j}) = P_{max}(s_{i,j,k}, C_i) - P_{min}(s_{i,j,k}, C_i) \tag{6}$$

$$CU(s_i) = \frac{y_{i,j} - P_{min}(s_{i,j,k}, C_i)}{P_{max}(s_{i,j,k}, C_i) - P_{min}(s_{i,j,k}, C_i)} \tag{7}$$

where P_{min} and P_{max} are respectively the two functions estimating the minimum and the maximum value of probability for the original class C_i predicted for the sentence s_i, over the new 100 sentence representation $s_{i,j,k}$; y_i is the sentence token $s_{i,j}$ original WMAL score. As stated by Fouladgar et al. (Fouladgar et al., 2020) we decided to use these values both in the explanation given to the end-user, graphically showing how much each sentence term is *"important"* and *"useful"* for the *prediction* obtained from the here proposed hybrid BERT-WMAL model.

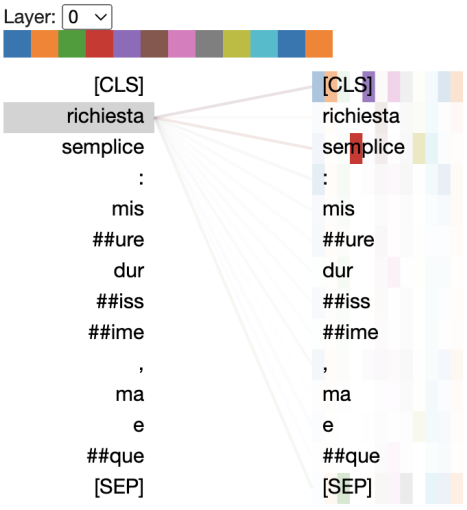


Figure 4: Example of multi-head attention mechanism. For each layer of the transformer, the attention value indicates how much a word on the left is responsible for the learning of the representation of each word on the right. In this example, 12 attention heads, computed in parallel independently from each other, are visualized.

Another strategy of explanation that we decided to implement is based on the scores of the different heads of attention of the BERT model. The BERT model is based on the Transformer architecture (Vaswani et al., 2017), which exploits the multi-head attention mechanism (MHA) as a strategy to differently focus on all the pairs of words composing the sentence in the model input. In BERT, we can find for each layer multiple attention heads. In particular, in a base BERT model, such as AlBERTo, we find 12 heads of attention for each of the 12 encoding layers. An attention head takes as input a sequence of embedding vectors $h = [h_1, ..., h_n]$ corresponding to the n tokens $e_i \in \mathbb{R}^n$ of the input sentence. In order to be correctly processed by the attention function, each vector h_i is transformed into a query, key, and value vectors q_i, k_i, v_i through separate linear transformations. By describing the attention mechanism from the perspective of a single token e_i attending to all input tokens the key vectors k_i are aggregated into the key matrix $K = [k_0, ..., k_n]$ and the value vectors v_i are aggregated into the value matrix $V = [v_0, ..., v_n]$. The attention vector a_i for the token e_i is then computed as in Eq. 8.

$$a_i = softmax\left(\frac{q_i^T \cdot K}{\sqrt{d_k}}\right) \tag{8}$$

where d_k is the length of a query or key vector, and softmax is then the standard softmax function (Goodfellow et al., 2016). The attention vector $a_i \in \mathbb{R}^n$ now contains an attention weight for each input token, including itself. Then this vector is multiplied by the value matrix V to compute the output of the self-attention operation. The outputs of all heads are then concatenated and fed through a linear layer to compute the output of the self-attention block for a single token. Figure 4 shows a visual example of a multi-head attention mechanism.

In our explanatory approach, for each token we used aggregate attention values to indicate how much attention each token receives in the classification process from the other neighborhoods composing the input sentence. Specifically, we have focused on the first layer of the model (Layer 0), which being the one closest to the original representation of the sentence, is the one most significant for obtaining a specific weight to each lexical token.
Focusing on a single attention head, if we consider $a_{i,j}$ the attention received from the token e_i to the token e_j, we calculate the following average function (Eq. 9):

$$rec\_att(e_j) = \frac{\sum_{i=0}^{n} a_{i,j}}{n} \tag{9}$$

Since only one aggregate attention value is obtained for each of the m attention heads in the model, it was decided to calculate an average again among the values obtained. The final score of our explanatory attention $expl\_att(e_j)$ of

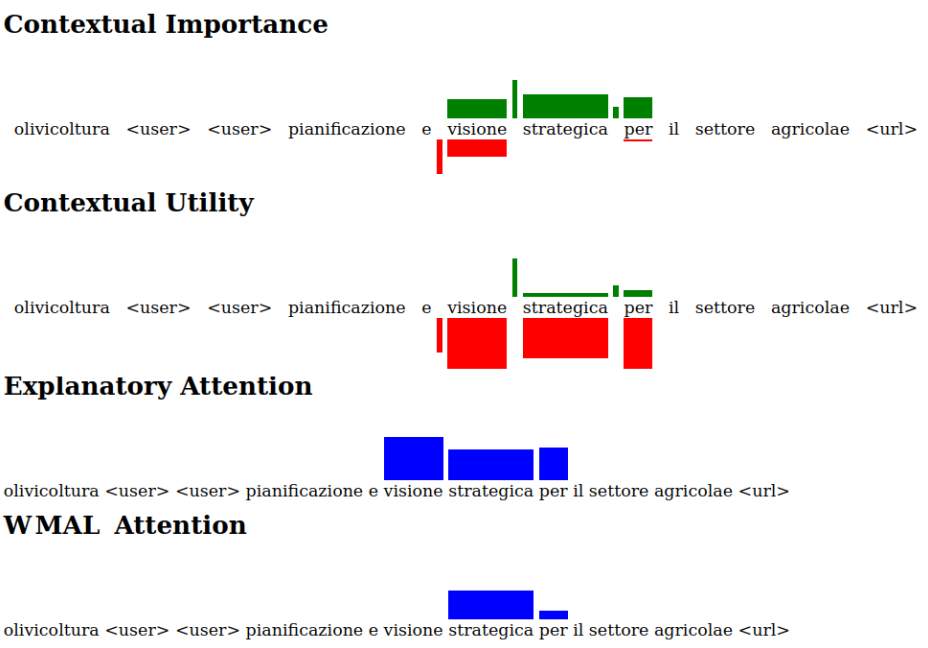


Figure 5: Visualization of importance, utility, attention and WMAL-attention for each word of the example tweet (translated from Italian: *Olive cultivation <user> <user> planning and strategic vision for the agriculture sector*). Utility and attention are computed for both the positive polarity (in green) and the negative polarity (in red).

the single token e_j is showed in Eq. 10.

$$expl\_att(e_j) = \frac{\sum_{i=0}^{m} rec\_att_m(e_j)}{m} \tag{10}$$

When individual words are split into multiple tokens by the BERT Wordpiece tokenizer, it was chosen to assign the highest explanatory attention value among those obtained from its components to the entire word. The final score obtained for each word in the component sentence was used through a graphical method to show an explanation of the classification process.
Finally, the last strategy we decided to implement combines the value of WMAL with that of explanatory attention. In particular, being available for each word of the input sentence the polarized score of the WMAL lexicon and that of explanatory attention both, we decided to multiply them as reported in Eq. 11.

$$wmal\_att(e_j) = expl\_att(e_j) \cdot WMAL(e_j) \tag{11}$$

Not only does this information give us a graphic value of a value of relevance of the term in the sentence, which is proportional to its emotional expressiveness, but it also guides us through its sign in the interpretation of the polarity assigned to it by the polarized lexicon. Figure 5 shows an example of how the word-level information can be used to assess the contribution given by each word toward the final prediction, for comparative purposes but also as a visual tool for interpreting the results of the model. In the figure, the bars on top of the words represent the method-specific scores. Contextual Importance and Contextual Utility have two scores, one for the positive polarity (green bar above the word) and one for the negative polarity (red bar below the word), while the two attention-based methods have only one score, represented by the blue bar above the words.

5.1. Evaluation and Discussion of Results

In order to test the validity of the explanation methods we have proposed, we performed an experiment involving human judges. We asked four raters, native speakers of Italian, to judge the quality of the explanations provided by three baseline methods plus our approach. More precisely, we selected a subset of AGRITREND made of 64 sentences, 32 with positive polarity gold labels and 32 with negative polarity gold labels. We then created four evaluation

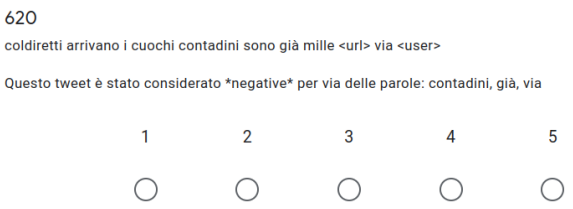


Figure 6: A screenshot of part of the human evaluation interface. The evaluators were shown a text identified by a number (620 in the example), the predicted label (negative), three words (*contadini, già, via)*) and a 1–5 scale to express to what extent the three words explain the prediction of the label on the text.

Table 8
Results of the human evaluation in terms of average score given by the human judges (on a scale 1–5) and their standard deviation.

Method	Polarity	Av. Score	St. Deviation	CV
Explanatory Attention	All	2.62	1.49	0.57
Contextual Importance		2.95	1.36	0.46
Contextual Utility		2.84	1.40	0.49
WMAL Attention		**3.20**	1.38	0.43
Explanatory Attention	Positive	2.78	1.53	0.55
Contextual Importance		2.78	1.23	0.44
Contextual Utility		2.59	1.34	0.52
WMAL Attention		**3.21**	1.18	0.37
Explanatory Attention	Negative	2.46	1.45	0.59
Contextual Importance		3.12	1.47	0.47
Contextual Utility		3.09	1.44	0.47
WMAL Attention		**3.18**	1.57	0.49

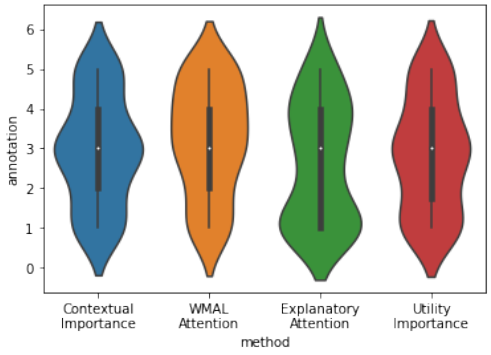
datasets so that 16 sentences (one-fourth of the set) were paired with the three top-ranked words of each method. The four evaluation sets contain different pairings, and their union contains all possible sentence-method combinations. Therefore, each of the four raters was prompted with 16 sentences for each of the four methods.

We created a form[7] for each of the raters, where, for each of the 64 sentences, the text, a sentiment polarity label, three explanatory words (without numerical scores) and a rating scale are shown. The label shown is the prediction given by one of the four methods. However, the method that produced the label is not shown to the rater. The question posed to the rater is, "How much these three words are responsible for giving the text the connotation indicated by the label?" The rating scale is a Likert scale with five discrete values ranging from 1 (the explanation is not related to the label) to 5 (the words perfectly explain the prediction). An example of the evaluation interface is shown in Figure 6.

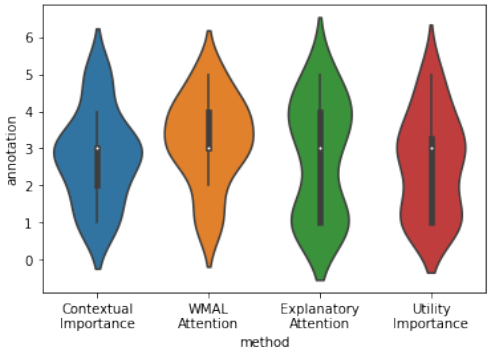
Table 8 summarizes the results of this human evaluation. The novel explanation method based on the affective lexicon led to the highest quality explanations overall, according to the human judges, and also when considering only the positive and negative instances (according to the gold standard label) separately. Interestingly, the attention mechanism alone produces the explanations judged as the worst by the human annotators, while importance and utility fare in the middle of the ranking. The explanations for the negative instances are judged on average to be of better quality than their positive counterparts, however, with a higher standard deviation.

The actual distributions of the scores given by the human judges, shown as violin plots in Figure 7, further highlighted how the attention-WMAL and the attention alone were found to be negatively skewed (i.e., rates more asymmetrical in the disagreement than in the agreement polarity) for negative instances and inflated around rate 3 for positive. Moreover, the average scores around the middle of the scale (3), were more representative for positive than negative instances (0.37 vs. 0.49 coefficient of variation (CV), in Table 8). This indicates a better human consensus on negative instances inferred by the attention-WMAL model with respect to the positive instances. This result is also reflected by the distributions of the simple attention model which, conversely, were found positively skewed (i.e., asymmetry in agreement rates) for negative instances and bimodal for positive.

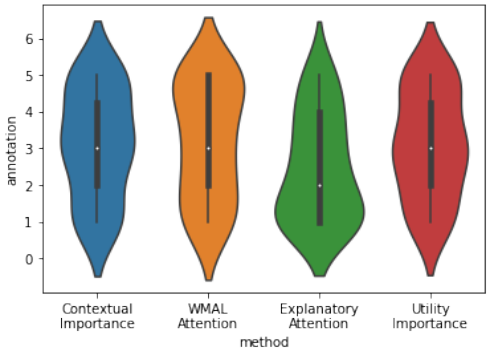
[7] We used Google Forms and its API to programmatically create the forms from the evaluation sets and show them to the raters.



(a) All instances



(b) Positive instances



(c) Negative instances

Figure 7: Violin plots of the distribution of human judgments across methods.

6. Implications of Research, Limits and Challenges

The SA model proposed in this work demonstrates how a hybrid approach based on neural networks and a polarized lexicon can achieve results comparable to the state of the art but with the added ability to explain the prediction to the end-user. We chose as our application domain the polarity detection for social media contents in a language different from English to enrich scientific research in natural language processing for minor languages. Nevertheless, the proposed approach can be generalized and usable with further transformer models and polarized lexicons, even for English or less widespread languages. The main limitation deals with the definition of a polarized lexicon through manual annotation operations. Not only is it a time-consuming automatic task, but it is often demanding and complex even for a human being. Therefore, in possible future applications, it could be relevant to investigate strategies for the automatic creation of polarized lexicons, such as the one proposed in (Bandhakavi, Wiratunga and Massie, 2021). After the operations of fine-tuning, the execution of the model in its totality can be carried out in real time. The explanation phase needs instead to execute a high number of predictions, 100 for each term occurring in the input sentence. This could be a limit for low-performing machines and can determine an excessive workload. This limitation presents a possible direction of improvement for real application in everyday settings. Research regarding explainable AI is ongoing, and the model proposed here actively contributes to this direction. In fact, it may represent one of the first feature ablation explanation approaches for hybrid models. It can be the basis for new work on the topic that could provide effective, human-like explanations that make the model transparent. The use of an explainable SA model can involve numerous applications in different contexts of use, including the industrial one, for instance, in recommender systems. Considering the category of content-based or hybrid approaches, it would be possible to create a user model based on opinions left by the user online. During the recommendation process, it will not only be possible to say that a product has been recommended because a specific review has been left online, but also how it has been evaluated and considered by the system with details on the polarized words identified as relevant for the result of the recommendation. Another application area could be the moderation of textual content on social media.

Although explicitly prohibited by many terms of service, negative and hate content is widespread on many social media platforms. One of the strategies most commonly implemented by social media providers for limiting this problem is the use of human moderators. Deciding whether a piece of content is inappropriate for the platform is normally a personal burden carried by individuals, who are often underpaid and work long hours a day with inappropriate content. This can have an impact on their personal lives and their mental health. Unfortunately, these efforts are often limited, and only a small percentage of inappropriate content is actually processed and therefore censored Chen, McKeever and Delany (2019). Consequently, with this work, we could promote a system capable of supporting moderators by providing a pretty detailed list of possible hate textual contents, that is, those identified as hate speech by our SA model. Moreover, we could support the human decision by also providing a human-like message to explain the decision taken by the platform on the basis of the polarized lexicon adopted. These possibilities demonstrate the broad applicability of the proposed approach by demonstrating its extensive impact on possible future research directions on the topic.

7. Conclusion and Future Work

The BERT-WMAL hybrid model proposed in this paper shows how to combine the power of an approach based on deep neural transformer architectures with that of a lexicon-based approach. In particular, we have shown how a hybrid architecture is able to perform better than single SA models, based on lexicon or on transformers only. On the one hand, the emotional lexicon has proved to play a crucial role in the model in order to obtain classification performances at least equal to those at the state-of-the-art. On the other hand, this aspect allowed us to show the impact of the emotional lexicon in terms of performance as well as to build a final model that featured a higher degree of explainability.
We proposed three different approaches with which it is possible to explain the model results, and, in particular, the one based on the linear combination of WMAL and attention levels proved to be the most promising according to the human evaluation we provided. Despite being demonstrated in the Italian language and on a polarity detection task, the proposed approach can be easily extended to other application domains and different languages. In fact, it is sufficient to model a neural network that obtains as input the contextual embeddings of the transformer network and a lexicon with a relevance value measured for the application domain. This makes the hybrid model presented here an excellent starting point for further research in the current trend that is focused on the explainable of AI.

In future work, we plan to evaluate the model with transformers of different types, more recent architectures, applicable to linguistic tasks other than SA. The WMAL scores can be computed on different language sources and additional neutrality thresholds. In general, however, we would like to apply the principle behind WMAL for lexicons associated with different linguistic tasks. The application of our method does not need to be restricted to social media either, although the abundance of accessible textual material from such a source makes it especially easy to apply a frequency-based ponderation scheme to an affective lexicon.

One aspect of (Vassallo et al., 2020) that was not fully exploited in this work is the imbalance of performance across polarity measured when adopting lexicon-based methods, and its relationship with the use of arbitrary or empirically determined thresholds for assigning polarity labels to text instances. In particular, the F1-score on the negative polarity is often higher than the one on the positive polarity class, i.e., the negative polarity of tweets is better predicted than the positive polarity. A possible explanation for this finding is that the vocabulary of negative sentiments is richer and more sought-after than that of positive sentiments. The results presented in this paper on the AGRITREND dataset show that this effect is amplified the more the topic is specific, an effect already measured in (Vassallo et al., 2020) and explained in terms of the use of less frequent words in topic-specific messages. We believe that this research direction, along with the ones mentioned earlier in these conclusions, will be able to make use of the results obtained here to address the dual problem of the efficiency of deep learning models and their ability to be transparent and explainable.

Acknowledgments

We would like to thank Dr. Antonella Di Fonzo of CREA, Research Centre for Agricultural Policies and Bioeconomy, for her support in setting up the AGRITREND dataset. The work of Marco Polignano has been supported by Apulia Region, Italy through the project ”Un Assistente Dialogante Intelligente per il Monitoraggio Remoto di Pazienti” (Grant n. 10AC8FB6) in the context of ”Research for Innovation - REFIN”.

CRediT authorship contribution statement

Marco Polignano: Methodology; Software; Writing - Original Draft. **Valerio Basile:** Conceptualization; Writing - Review & Editing. **Pierpaolo Basile:** Validation; Writing - Review & Editing. **Giuliano Gabrieli:** Investigation; Data Curation; Writing - Review & Editing. **Marco Vassallo:** Investigation; Data Curation; Writing - Review & Editing. **Cristina Bosco:** Resources; Writing - Review & Editing.

References

Agüero-Torales, M.M., Salas, J.I.A., López-Herrera, A.G., 2021. Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing 107, 107373.

van Aken, B., Winter, B., Löser, A., Gers, F.A., 2019. How does bert answer questions? a layer-wise analysis of transformer representations, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1823–1832.

Angelov, P., Soares, E., 2020. Towards explainable deep neural networks (xdnn). Neural Networks 130, 185–194.

Arous, I., Dolamic, L., Yang, J., Bhardwaj, A., Cuccu, G., Cudré-Mauroux, P., 2021. Marta: Leveraging human rationales for explainable text classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5868–5876.

Baccianella, S., Esuli, A., Sebastiani, F., 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., in: Lrec, pp. 2200–2204.

Bacco, L., Cimino, A., Dell'Orletta, F., Merone, M., 2021a. Extractive summarization for explainable sentiment analysis using transformers, in: Abbès, S.B., Hantach, R., Calvez, P., Buscaldi, D., Dessì, D., Dragoni, M., Recupero, D.R., Sack, H. (Eds.), Joint Proceedings of the 2nd International Workshop on Deep Learning meets Ontologies and Natural Language Processing (DeepOntoNLP 2021) & 6th International Workshop on Explainable Sentiment Mining and Emotion Detection (X-SENTIMENT 2021) co-located with co-located with 18th Extended Semantic Web Conference 2021, Hersonissos, Greece, June 6th - 7th, 2021 (moved online), CEUR-WS.org. pp. 62–73. URL: http://ceur-ws.org/Vol-2918/paper5.pdf.

Bacco, L., Cimino, A., Dell'Orletta, F., Merone, M., 2021b. Explainable sentiment analysis: A hierarchical transformer-based extractive summarization approach. Electronics 10. URL: https://www.mdpi.com/2079-9292/10/18/2195, doi:10.3390/electronics10182195.

Bandhakavi, A., Wiratunga, N., Massie, S., 2021. Emotion-aware polarity lexicons for twitter sentiment analysis. Expert Systems 38, e12332.

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V., 2016. Overview of the evalita 2016 sentiment polarity classification task, in: Basile, P., Corazza, A., Cutugno, F., Montemagni, S., Nissim, M., Patti, V., Semeraro, G., Sprugnoli, R. (Eds.), Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016, CEUR-WS.org. URL: http://ceur-ws.org/Vol-1749/paper_026.pdf.

Basile, P., Croce, D., Basile, V., Polignano, M., 2018. Overview of the evalita 2018 aspect-based sentiment analysis task (absita), in: EVALITA Evaluation of NLP and Speech Tools for Italian, CEUR. pp. 1–10.

Basile, V., Nissim, M., 2013a. Sentiment analysis on Italian tweets, in: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Atlanta, Georgia. pp. 100–107. URL: https://aclanthology.org/W13-1614.

Basile, V., Nissim, M., 2013b. Sentiment analysis on italian tweets, in: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 100–107.

Baziotis, C., Pelekis, N., Doulkeridis, C., 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada. pp. 747–754.

Birjali, M., Kasri, M., Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems 226, 107134. URL: https://www.sciencedirect.com/science/article/pii/S095070512100397X, doi:https://doi.org/10.1016/j.knosys.2021.107134.

Bodria, F., Panisson, A., Perotti, A., Piaggesi, S., 2020. Explainability methods for natural language processing: Applications to sentiment analysis, in: Agosti, M., Atzori, M., Ciaccia, P., Tanca, L. (Eds.), Proceedings of the 28th Italian Symposium on Advanced Database Systems, Villasimius, Sud Sardegna, Italy (virtual due to Covid-19 pandemic), June 21-24, 2020, CEUR-WS.org. pp. 100–107. URL: http://ceur-ws.org/Vol-2646/18-paper.pdf.

Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317.

Cambria, E., Liu, Q., Decherchi, S., Xing, F., Kwok, K., 2022. Senticnet 7: a commonsense-based neurosymbolic ai framework for explainable sentiment analysis. Proceedings of LREC 2022 .

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R., 2018. Universal sentence encoder for english, in: Blanco, E., Lu, W. (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics. pp. 169–174. URL: https://doi.org/10.18653/v1/d18-2029, doi:10.18653/v1/d18-2029.

Chen, H., McKeever, S., Delany, S.J., 2019. The use of deep learning distributed representations in the identification of abusive text, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 125–133.

Clos, J., Wiratunga, N., Massie, S., 2017. Towards explainable text classification by jointly learning lexicon and modifier terms, in: IJCAI-17 Workshop on Explainable AI (XAI), p. 19.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (Eds.), Proceedings of the

58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics. pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747, doi:10.18653/v1/2020.acl-main.747.

Deshmane, A.A., Friedrichs, J., 2017. TSA-INF at SemEval-2017 Task 4: An ensemble of deep learning architectures including lexicon features for Twitter sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 802–806.

Devlin, J., Chang, M., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding, in: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics. pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423, doi:10.18653/v1/n19-1423.

Dridi, A., Reforgiato Recupero, D., 2019. Leveraging semantics for sentiment polarity detection in social media. International Journal of Machine Learning and Cybernetics 10, 2045–2055.

El-Din, D.M., 2016. Enhancement bag-of-words model for solving the challenges of sentiment analysis. International Journal of Advanced Computer Science and Applications 7.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S., 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: Palmer, M., Hwa, R., Riedel, S. (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics. pp. 1615–1625. URL: https://doi.org/10.18653/v1/d17-1169, doi:10.18653/v1/d17-1169.

Fiok, K., Karwowski, W., Gutierrez, E., Wilamowski, M., 2021. Analysis of sentiment in tweets addressed to a single domain-specific twitter account: Comparison of model performance and explainability of predictions. Expert Systems with Applications 186, 115771. URL: https://www.sciencedirect.com/science/article/pii/S0957417421011428, doi:https://doi.org/10.1016/j.eswa.2021.115771.

Fouladgar, N., Alirezaie, M., Främling, K., 2020. Decision explanation: applying contextual importance and contextual utility in affect detection, in: Italian Workshop on Explainable Artificial Intelligence, XAI. it 2020, co-located with 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2020), Online Event, November 25-26, 2020, Technical University of Aachen. pp. 1–13.

Främling, K., 2020. Decision theory meets explainable ai, in: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer. pp. 57–74.

Giatsoglou, M., Vozalis, M.G., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C., 2017. Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications 69, 214–224.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Gupta, I., Joshi, N., 2020. Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. J. Intell. Syst. 29, 1611–1625. URL: https://doi.org/10.1515/jisys-2019-0106, doi:10.1515/jisys-2019-0106.

van Heuven, W., Mandera, P., Keuleers, E., Brysbaert, M., 2014. SUBTLEX-UK: a new and improved word frequency database for British English. Quarter Journal of Experimental Psychology 67, 1176–90.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Jurek, A., Mulvenna, M.D., Bi, Y., 2015. Improved lexicon-based sentiment analysis for social media analytics. Secur. Informatics 4, 9. URL: https://doi.org/10.1186/s13388-015-0024-x, doi:10.1186/s13388-015-0024-x.

Koufakou, A., Pamungkas, E.W., Basile, V., Patti, V., 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online. pp. 34–43. URL: https://aclanthology.org/2020.alw-1.5, doi:10.18653/v1/2020.alw-1.5.

Kudo, T., Richardson, J., 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Blanco, E., Lu, W. (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics. pp. 66–71. URL: https://doi.org/10.18653/v1/d18-2012, doi:10.18653/v1/d18-2012.

Lahase, A.R., Shelke, M., Jagdale, R., Deshmukh, S., 2022. A survey on sentiment lexicon creation and analysis, in: IOT with Smart Systems. Springer, pp. 579–587.

Li, W., Zhu, L., Shi, Y., Guo, K., Cambria, E., 2020. User reviews: Sentiment analysis using lexicon integrated two-channel cnn-stm family models. Applied Soft Computing 94, 106435.

Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5, 1–167.

Liu, B., Zhang, L., 2012. A survey of opinion mining and sentiment analysis, in: Mining text data. Springer, pp. 415–463.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR abs/1907.11692. URL: http://arxiv.org/abs/1907.11692, arXiv:1907.11692.

London, A.J., 2019. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report 49, 15–21.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777.

Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C., 2011. Learning word vectors for sentiment analysis, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pp. 142–150.

Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Vossen, P., Azpeitia, A., 2014. Generating polarity lexicons with WordNet propagation in 5 languages, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland. pp. 1155–1161.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A., 2018. Advances in pre-training distributed word representations, in: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/721.html.

Mohammad, S., Kiritchenko, S., 2018. Wikiart emotions: An annotated dataset of emotions evoked by art, in: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/966.html.

Nayak, A., Natarajan, D., 2016. Comparative study of naive Bayes, support vector machine and random forest classifiers in sentiment analysis of Twitter feeds. International Journal of Advance Studies in Computer Science and Engineering (IJASCSE) 5, 14–17.

Okanohara, D., Tsujii, J., 2005. Assigning polarity scores to reviews using machine learning techniques, in: International Conference on Natural Language Processing, Springer. pp. 314–325.

Pamungkas, E.W., Patti, V., 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy. pp. 363–370. URL: https://aclanthology.org/P19-2051, doi:10.18653/v1/P19-2051.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Perikos, I., Kardakis, S., Hatzilygeroudis, I., 2021. Sentiment analysis using novel and interpretable architectures of Hidden Markov Models. Knowledge-Based Systems 229, 107332. URL: https://www.sciencedirect.com/science/article/pii/S0950705121005943, doi:https://doi.org/10.1016/j.knosys.2021.107332.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations, in: Walker, M.A., Ji, H., Stent, A. (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics. pp. 2227–2237. URL: https://doi.org/10.18653/v1/n18-1202, doi:10.18653/v1/n18-1202.

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., Basile, V., 2019a. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets, in: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, CEUR. pp. 1–6.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., 2019b. A comparison of word-embeddings in emotion detection from text using bilstm, cnn and self-attention, in: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 63–68.

Polignano, M., Basile, V., Basile, P., de Gemmis, M., Semeraro, G., 2019c. Alberto: Modeling italian social media language with bert. IJCoL. Italian Journal of Computational Linguistics 5, 11–31.

Pota, M., Ventura, M., Fujita, H., Esposito, M., 2021. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. Expert Systems with Applications 181, 115119.

Qazi, A., Raj, R.G., Hardaker, G., Standing, C., 2017. A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges. Internet Research .

Ras, G., van Gerven, M., Haselager, P., 2018. Explanation methods in deep learning: Users, values, concerns and challenges, in: Explainable and interpretable models in computer vision and machine learning. Springer, pp. 19–36.

Raschka, S., 2014. An overview of general performance metrics of binary classifier systems. CoRR abs/1410.5330. URL: http://arxiv.org/abs/1410.5330, arXiv:1410.5330.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Rosenthal, S., Farra, N., Nakov, P., 2017. Semeval-2017 task 4: Sentiment analysis in Twitter, in: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), pp. 502–518.

Samek, W., Müller, K.R., 2019. Towards explainable artificial intelligence, in: Explainable AI: interpreting, explaining and visualizing deep learning. Springer, pp. 5–22.

Samek, W., Wiegand, T., Müller, K., 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. CoRR abs/1708.08296. URL: http://arxiv.org/abs/1708.08296, arXiv:1708.08296.

Seema, S., et al., 2022. Deep learning approaches for sentiment analysis challenges and future issues, in: Deep Learning Applications for Cyber-Physical Systems. IGI Global, pp. 27–50.

Sharma, R., Nigam, S., Jain, R., 2014. Polarity detection at sentence level. International journal of computer applications 86.

Silveira, T.D.S., Uszkoreit, H., Ai, R., 2019. Using aspect-based analysis for explainable sentiment predictions, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer. pp. 617–627.

Strapparava, C., Valitutti, A., et al., 2004. Wordnet affect: an affective extension of Wordnet., in: Lrec, Lisbon. p. 40.

Tamburini, F., 2020. How "bertology" changed the state-of-the-art also for italian NLP, in: Monti, J., Dell'Orletta, F., Tamburini, F. (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, CEUR-WS.org. URL: http://ceur-ws.org/Vol-2769/paper_79.pdf.

Tiddi, I., Schlobach, S., 2022. Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence 302, 103627.

Vassallo, M., Gabrieli, G., Basile, V., Bosco, C., 2019. The tenuousness of lemmatization in lexicon-based sentiment analysis, in: Bernardi, R., Navigli, R., Semeraro, G. (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, CEUR-WS.org. URL: http://ceur-ws.org/Vol-2481/paper74.pdf.

Vassallo, M., Gabrieli, G., Basile, V., Bosco, C., 2020. Polarity imbalance in lexicon-based sentiment analysis, in: Monti, J., Dell'Orletta, F., Tamburini, F. (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, CEUR-WS.org. URL: http://ceur-ws.org/Vol-2769/paper_36.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Vig, J., 2019. Bertviz: A tool for visualizing multihead self-attention in the bert model, in: ICLR Workshop: Debugging Machine Learning Models.

Vilone, G., Longo, L., 2020. Explainable artificial intelligence: a systematic review. CoRR abs/2006.00093. URL: https://arxiv.org/abs/2006.00093, arXiv:2006.00093.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Linzen, T., Chrupala, G., Alishahi, A. (Eds.), Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018, Association for Computational Linguistics. pp. 353–355. URL: https://doi.org/10.18653/v1/w18-5446, doi:10.18653/v1/w18-5446.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144. URL: http://arxiv.org/abs/1609.08144, arXiv:1609.08144.

Zacchetta, E., Baroni, M., 2006. Morph-it! A free corpus-based morphological resource for the Italian language, in: Proceedings of Corpus Linguistics 2005.

Zhang, Y., Chen, X., et al., 2020. Explainable recommendation: A survey and new perspectives. Foundations and Trends® in Information Retrieval 14, 1–101.

Zhao, A., Yu, Y., 2021. Knowledge-enabled bert for aspect-based sentiment analysis. Knowledge-Based Systems 227, 107220. URL: https://www.sciencedirect.com/science/article/pii/S0950705121004822, doi:https://doi.org/10.1016/j.knosys.2021.107220.

Zheng, G., Mukherjee, S., Dong, X.L., Li, F., 2018. Opentag: Open attribute value extraction from product profiles, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1049–1058.

Zipf, G.K., 1949. Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology. Addison-Wesley.