

Deceptive reviews and sentiment polarity: Effective link by exploiting BERT[☆]Rosario Catelli^{a,*}, Hamido Fujita^{b,c,d}, Giuseppe De Pietro^a, Massimo Esposito^a^a Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), Naples, Italy^b Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Viet Nam^c National Taipei University of Technology, Taipei, Taiwan^d i-somet Incorporation association, Morioka, Japan

ARTICLE INFO

Keywords:

Deceptive reviews
Sentiment
BERT
Multi-label
Deep learning
Neural language model

ABSTRACT

Today, reviews are the advertising medium par excellence through which companies are able to influence customers' spending decisions. Although the initial purpose of reviews was to provide companies with a feedback tool to improve products and services based on customer needs, they soon became a way to climb the sales rankings, often illegally. In fact, deceptive and fake reviews have managed to evade the often non-existent means of validation of online platforms, proliferating a new business. To combat this phenomenon, several classification methods have been developed to train automated tools in the arduous task of distinguishing between genuine and misleading reviews, the most recent based on machine and deep learning techniques. This paper proposes a multi-label classification methodology based on the Google BERT neural language model to build a deceptive review detector aided by its sentiment awareness: improved modeling of the link between sentiment polarity and deceptiveness during the fine-tuning phase by exploiting the Binary Cross Entropy with Logits loss function adds to the advantages provided by pre-trained contextual models, which are able to capture word polysemy through word embeddings and benefit from pre-training on huge corpora. Tests were performed on the Deceptive Opinion Spam Corpus and Yelp New York City datasets, providing a quantitative and qualitative analysis of the results which, when compared with the state of the art available in the literature, showed an encouraging increase in performance.

1. Introduction

The advent and mass dissemination of social media has led to a strong proliferation of users' opinions in a wide range of areas such as news and products, previously the preserve of only experts in the field who were widely visible through traditional media such as TV and newspapers. At the same time, the progressive increase in online sales of products and services for which it is necessary to pay before being able to touch or try the goods purchased has meant that the importance of reviews as a tool, also for advertising, to influence sales performance has increased exponentially: nowadays, according to several studies,¹ more than 9 out of 10 consumers rely on reviews before making purchases. Moreover, this business practice has spread rapidly in both the private and public sectors. Trusting reviews before deciding whether to buy a product on e-commerce sites has become a common ritual, as well as reading what other users say before paying

to listen to or watch streaming media content, or deciding which hotel to stay in and which travel destination to choose. Similarly, reading reviews to find out which cultural heritage sites to divert money to during a leisure trip has become the practice for the most attentive and regular consumers and travelers: museums, archaeological sites, green and water parks, nature reserves and so on have become the subject of careful planning and management of economic resources also by local authorities, which seek through favorable feedback from visitors to trigger positive circles for the surrounding economy, such as transport, reception and restoration.

For these reasons, the most important parameters have turned out to be both the volume and the score of reviews (Maslowska et al., 2017), since according to the most recent estimates nine out of ten consumers document themselves through this tool before making a choice (Kumar et al., 2018) and, as in a vicious circle, the more reviews the greater the possibility that the service is purchased because it is considered

[☆] This work is supported by the Innovation for Data Elaboration in Heritage Areas (IDEHA) project which has received funding from the National Operational Programme (PON) of the Italian Ministry of Education, University and Research (MIUR): code ARS01_00421 (Decree n.2059, 02 August 2018).

* Corresponding author.

E-mail addresses: rosario.catelli@icar.cnr.it (R. Catelli), h.fujita@hutech.edu.vn, h.fujita-799@acm.org (H. Fujita), giuseppe.depietro@icar.cnr.it (G. De Pietro), massimo.esposito@icar.cnr.it (M. Esposito).

¹ <https://www.qualtrics.com/blog/online-review-stats/>.

well-known and broken-in, the more so as it turns out to be expensive (Askalidis & Malthouse, 2016). At the same time, the problem of misleading reviews has arisen, capable of demonizing or sanctifying products and services, thus pushing them commercially towards failure or success: it is the so-called *crowd-turfing* phenomenon (Lee et al., 2014). Similarly, this activity has repercussions on commercial services operating in the most disparate sectors, e.g. agribusiness, restaurants, healthcare, all the more so as these services are included within specialized search engines that rank them through proprietary algorithms (Mukherjee et al., 2013; Santosh & Mukherjee, 2016) also with the help of reviews, such as TripAdvisor or Yelp. From the perspective of the scientific literature, the problem of deceptive reviews has generally been addressed as a spam and deception detection problem. Technically, binary classification tools coupled with text features have been employed, like word and part-of-speech n-gram or structural features obtained from syntactic parsing (Feng, Banerjee et al., 2012) or psycho-linguistic features obtained using *Linguistic Inquiry and Word Count* (Hernández-Castañeda et al., 2017; Ott et al., 2013, 2011; Pennebaker et al., 2015), *Latent Dirichlet Allocation* topics modeling (Hernández-Castañeda et al., 2017) or non-verbal features related to reviewer behavior (Aghakhani et al., 2018; Stanton & Irissappane, 2019; You et al., 2018) and diverse techniques for complex arising problems such as that of sock-puppets (Hosseinia & Mukherjee, 2017). Recently, techniques based on deep learning have become widespread to automatically solve problems such as feature engineering, which have become increasingly complex over time, and to better represent texts by taking context into account (Akbik et al., 2018; Devlin et al., 2019; Peters et al., 2018). Specifically with regard to the detection of deceptive reviews, Aghakhani et al. (2018) tried to tackle the problem using GAN networks, with a generator and two discriminators.

Several recent studies have shown how sentiment is related in a particular way to misleading opinions: for example Zaeem et al. (2020) have shown the presence of a statistical correlation between real opinions and positive sentiment or false opinions and negative sentiment, going to confirm empirically the presence of psychological aspects related to sentiment and the attempt of the author to detach himself all the more when his writing is negative and false. Recently Shehnepoor et al. (2020) built ScoreGAN, a fraud reviews detector, that outperformed the previous FakeGAN (Aghakhani et al., 2018) combining it with NetSpam (Shehnepoor et al., 2017), using Regulated GAN: their comparative paper showed that score-based features are the best at detecting opinion spam due to the fact that the fraudster tries to promote or denigrate products and services by operating on the score and sentiment of the review. Hence, methods aimed at identifying fake reviews should not ignore information about sentiment polarity when present in the available dataset, as in the case of the crowd-sourced *Deceptive Opinion Spam Corpus v1.4* (DOSC) (Ott et al., 2013, 2011), or retrievable from ratings, as in the case of the *Yelp New York City* dataset (YelpNYC) (Rayana & Akoglu, 2015). Furthermore, to the best of our knowledge, the Google BERT neural language model (Devlin et al., 2019), that has distinguished itself in recent years for achieving state-of-the-art performance in several tasks, is recently used for fake news detection (Ding et al., 2020; Kaliyar et al., 2021; Zhang et al., 2020) but not much for fake reviews detection (Kennedy et al., 2019; Kim et al., 2021), that is a different task (Zhou & Zafarani, 2020) because it can attract both malicious and normal users (Shao et al., 2018). Furthermore, BERT and variants have been widely used in multi-label and multi-class contexts (Cai et al., 2020; Tang et al., 2020; Tao & Fang, 2020), although not related to the recognition of deceptive reviews.

In the scientific literature it is possible to find works, such as Kennedy et al. (2019) and Kim et al. (2021), that exploit contextual language models pre-trained on large corpora and based on deep neural networks such as BERT to perform deceptive review recognition, or it is possible to find works, such as the one by Martínez-Torres and Toral (2019), that instead demonstrate how deceptive review recognition employing classical classifiers finds benefit from matching sentiment

polarity information by appropriately engineering it as features. As far as known, no one to date has tried to combine the advantages of these two approaches: it is in this direction that our contribution is oriented. In particular, the proposal of this article consists of the following contributions:

- the development of a deceptive review detection system based on pre-trained contextual language models, in particular by testing the performance of the BERT and DistilBERT models in the cased and uncased variants;
- the incorporation into the aforementioned system of sentiment polarity classification by exploiting the ability of the cited language models to capture the polysemy of words through word embeddings and effectively represent syntactic-semantic links also by detecting synonyms and antonyms often present in reviews such as fantastic, great, horrible, terrible, and so on;
- the modeling of a sentiment-aware deceptive review detection system by means of a multi-label classification scheme based on a Label Powerset (LP) technique that employs a tuple [deceptiveness, sentiment polarity] as powerset: in this way the deep neural networks that constitute language models are leveraged during the fine-tuning phase to automatically extract the advantageous hidden features for recognizing the deceptiveness of reviews also through sentiment information;
- the experimentation, the analysis and the comparison with the state of the art available in literature of the results obtained on a dataset both of small size, that is the DOSC dataset, and of medium-large size, that is the YelpNYC dataset.

In addition and differently from previous studies for the classification of misleading reviews (Kennedy et al., 2019; Kim et al., 2021), recent works regarding the instability of the fine-tuning process (Mosbach et al., 2021; Zhang et al., 2021) have been taken into account, suggesting a different approach from what was recommended by the creators of BERT, i.e. the use of 2–4 epochs regardless of the task and the dataset used. In particular it was empirically verified that, holding equal all other hyper-parameters generally suggested by the BERT authors and related to the classification of misleading reviews with both DOSC and YelpNYC datasets, the best solution falls outside the suggested range 2–4 but is not too far off (10, 15 or 20 epochs led to worse results than 5 epochs).

The remainder of this paper is structured as follows. In Section 2 backgrounds on the topic and the most important related works are illustrated. In Section 3 the dataset used and the architecture proposed are described. In Section 4 the experimental setup and the outcome are analyzed and discussed. Finally, in Section 5, some conclusions and future works are drawn.

2. Background and related works

In this section the scientific articles that form the foundation of this paper have been organized. In particular, an overview of the systems for detecting fake reviews is provided in Section 2.1, with an in-depth look at sentiment-aware systems in Section 2.1.1. In Section 2.2, a precise framing of the type of classification process examined was provided, namely that of multi-label classification. Finally, an outline of neural language modeling methods is provided in Section 2.3.

2.1. Fake review detection systems

The scientific literature (Jindal & Liu, 2008) identifies three main categories of review spam:

- untruthful opinions: reviews that are written deliberately to mislead readers and systems, positive (hyper spam) for promotional purposes or negative (defamatory spam) to damage reputation;

- reviews on brands only: reviews that are not specific to the product that should be the subject of analysis but are generic and related to brands, manufacturers or sellers;
- non-reviews: reviews that have advertising or irrelevant content and therefore do not contain opinions.

The latter two types are generally easy to detect thanks to content analysis and traditional learning and classification methods, whereas the first type is the one that still attracts most scientific research today because of its greater difficulty in detection: spam due to an untruthful opinion created ad-hoc is at first sight unmistakable from real and honest reviews. Moreover, due to the economic business behind this phenomenon, spammers are difficult to detect and consequently there is a gap in reliable data that can be used for training systems.

Among the most effective methods of detecting fake reviews are certainly those that employ machine learning techniques: on the many datasets available (like those based on Yelp, Amazon or TripAdvisor reviews) countless approaches have been tested, through machine learning based on both traditional statistics (such as Naive Bayes, Random Forest, Support Vector Machines or Ensemble) and deep neural networks (Convolutional Neural Networks, Recurrent Neural Networks or Generative Adversarial Network). A complete overview on this is given by Mohawesh et al. (2021), while in the following a brief discussion.

Machine learning techniques, whether supervised, semi-supervised or unsupervised (i.e. with high, low or no amounts of labeled training data respectively), are generally divided between review-centric or reviewer-centric or a combination of the two (Crawford et al., 2015). The former employ features contained in the individual review, such as linguistic, POS tagging, N-grams, textual, sentimental, rating or quality features, and thus entail a number of limitations: (1) spammers can evade anti-plagiarism checks by rewriting the texts of already known reviews, (2) such reviews are specialized according to content type, and (3) require real hand-labeled datasets. On the other hand, techniques that analyze the characteristics of reviewers, look at the frequency of reviews made by reviewers, the polarity (always negative? always positive?), the average length of texts, and so on: as Crawford et al. (2015) already noted, it is possible to identify threshold values by which false and true reviews can be distinguished on the basis of these parameters.

In this paper, it was chosen to focus on the strand of review-centric and not reviewer-centric approaches: rather than resorting to the use of features external to the reviews themselves, i.e. reviewers' social profiles that are not always available, it was preferred to dwell on the possibility of extracting additional information content from the reviews themselves by managing the polarity of sentiment with language models based on deep neural networks such as BERT.

Machine learning based approaches. These approaches can be supervised, semi-supervised and unsupervised.

Supervised learning techniques initially exploited duplicate reviews or copies of rather similar reviews as a basis for constituting labeled examples of misleading reviews (Jindal & Liu, 2008), a practice criticized for lack of reliability (Li et al., 2011; Ott et al., 2011) and superseded by pseudo-fake reviews obtained by paying anonymous workers (Ott et al., 2013, 2011): also this practice was criticized for the lack of *veracity* of the deceptive reviews due to authors' lack of domain knowledge, psychological state and experience in writing deceptive reviews (Mukherjee et al., 2013). Several authors found interesting results using Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM) or hybrids (Li et al., 2011) and proposing novel approaches (Li et al., 2013) or taking into account additive features (Banerjee et al., 2015). Related to the YelpNYC dataset, Khalifa et al. (2021) proposed an approach based on K-Nearest Neighbors classifier (K-NN) to extract additional features, balancing the dataset only in relation to the deceptiveness of reviews and splitting it 70/30 between training and testing, while Budhi et al. (2021) used LR, SVM

and Multilayer Perceptron (MLP) obtaining unsatisfying results due to the unbalancing of the dataset, that needed to be addressed.

Semi-supervised techniques started from realizing how the use of unexpected rules unmasked suspicious behavior of reviewers, and suggested exploiting such unexpected deviations from expected behavior (Jindal et al., 2010), hence modeling these aspects as variables (Feng, Xing et al., 2012).

Unsupervised models were developed due to the difficulty in labeling large amounts of data, exploiting text mining and improving random removal mechanisms of similar reviews that instead introduced a certain amount of distortion (Wu et al., 2010). Fei et al. (2013) introduced the Loopy Belief Propagation method to identify spammers, while Mukherjee et al. (2013) reformulated the problem as a Bayesian one defining the Author Spamicity Model.

Overall, supervised methods are clearly the most frequently reported in the literature, both classical (e.g., k-NN, LR, SVM, RF, GB and MLP) and newer, higher-performance ones based on language models employing deep neural networks (e.g., BERT and derivatives): the purpose of this paper is to understand whether these latter benefit further from integrating information content related to sentiment polarity into the deceptive review recognition task.

2.1.1. Sentiment-aware fake review detection systems

To date, a number of systems have been tested to detect fake reviews, many based on classical approaches employing machine learning techniques such as LR, SVM, or NB methods, and in several cases there was a question of how sentiment could be linked to the fakes, trying to take advantage of it in some way.

The importance of sentiment as a component of analysis then detection of deceptive reviews gradually emerged over time, highlighting how sentiment tended to be exaggerated in deceptive reviews, regardless of polarity, with some peculiar characteristics. For example, the use of first-person singular pronouns, e.g. *I*, *me*, *my*, should be carefully studied: from a psychological point of view, a deficiency of it leads to think that it is a way for the author to distance himself from the false and negative statements written from which consequently transpires a possible lack of personal experience (Ott et al., 2013), while on the other hand an overabundance of it would indicate that the deceivers try to emphasize their physical presence assuming that this increases their credibility (Li et al., 2014). This was also supported by Harris (2012), stating that deceptive reviews, in addition to being less readable than truthful ones, were more polarized and sentimental.

Li et al. (2011) integrated, among the different features of their spam review detection system, those based on sentiment by distinguishing between subjective or objective reviews (working at both word and sentence level) and positive or negative: in synergy with other features, the results showed improvements. Feng, Banerjee et al. (2012), stating that previous studies relied only on shallow lexicon-syntactic patterns, investigated syntactic stylometry for deception detection demonstrating that features driven from Context Free Grammar parse trees improved the recognition performance. Banerjee and Chua (2014) constructed a theoretical framework to distinguish bogus reviews from genuine reviews from three textual characteristics, comprehensibility, informativeness, and writing style: the differences between genuine and manipulated reviews tended to be more pronounced from the standpoint of comprehensibility and informativeness for reviews with negative sentiment, from the standpoint of writing style instead for reviews with positive sentiment. In a further development, Banerjee et al. (2015) took into account an additional element for classification purposes, namely the so-called cognitive indicators, showing positive and negative sentiment words as differentiating elements between original and fake reviews. Instead, Cagnina and Rosso (2015) proposed the use of different features for a low-dimensional representation of opinions, incorporating features into a SVMs classifier and obtaining competitive results while, investigating whether the use of emotions could help discriminate between true and misleading opinions, no encouraging

Table 1
Deceptiveness-Sentiment polarity relationships.

Authors	Mechanisms observed
Ott et al. (2013)	Use of first-person singular pronouns: deficiency to distance yourself (the deceiver) from negative feelings and falsehoods
Li et al. (2014)	Use of first-person singular pronouns: overabundance to emphasize physical presence hence deceivers' credibility
Harris (2012)	Lower readability and greater sentimental bias and emphasis
Li et al. (2011)	Distinction between subjectivity and objectivity, at the word and sentence level
Feng, Banerjee et al. (2012)	Syntactic stylometry: features driven from Context Free Grammar (CFG) parse trees
Banerjee and Chua (2014)	Comprehensibility, informativeness, and writing style
Banerjee et al. (2015)	Cognitive indicators
Cagnina and Rosso (2015)	Features integrated within SVM classifiers to represent emotion without encouraging results
Molla et al. (2017)	General sentiment classifier plus negative deception detector
Martinez-Torres and Toral (2019)	Content analysis approach based on unique attribute sets and sentiment orientation to obtain non-biased classifiers
Zaeem et al. (2020)	Statistical relationship between (a) positive sentiment and real news and (b) between negative sentiment and fake news
Hegde et al. (2021)	External sentiment dictionaries to support a hybrid ensemble model
Shang et al. (2021)	Opinion triplet extraction to add further labeling to target dataset and extract additional features to modify BERT embedding layer

results were obtained. Molla et al. (2017) proposed a system consisting of two main components, a general sentiment classifier and a negative deception detector, focusing their work only on reviews that had negative sentiment. Martinez-Torres and Toral (2019), following a content analysis approach based on both unique attribute sets and the sentiment orientation of reviews, stated that exploiting these aspects can yield non-biased classifiers. Zaeem et al. (2020) observed a statistically significant relationship between positive sentiment and real news and between negative sentiment and fake news. Hegde et al. (2021) leveraged a hybrid ensemble model based on both machine-learning and deep-learning classifiers, relying on external vocabularies to provide a rating relative to the sentiment of each review; instead, Shang et al. (2021) tried to leverage BERT, modifying its embeddings through additional sentiment information derived through further labeling of reviews provided by an external model. In both of the latter two cases, the experiments were conducted on the YelpNYC dataset but the dataset imbalance issues were not taken into account and the information regarding the split between training and testing of the dataset is lacking. A brief summary is reported in Table 1.

Until today, although there have been numerous proposals for the management and integration of the sentiment polarity within different classification systems of misleading reviews in order to take advantage of it, up to our knowledge no one has attempted such integration with the most modern and performing systems based on deep neural networks, i.e. language models such as BERT and derivatives: in this paper it is proposed to proceed to this integration modeling the problem as a multi-label classifier of LP type then exploit in this way the ability of automatic extraction of useful hidden features for the recognition of deceptive reviews taking advantage also of the contribution of their sentiment polarity.

2.2. Multi-label classification approaches

In recent years, classification systems have increasingly improved their performance, moving from rudimentary rule-based methods to machine then deep learning: the demand for multi-label classification systems has increased driven by market demands, e.g. Netflix or Spotify classifying movies and music with multiple different labels. In single-label classification systems, each example is associated with a single label l , with $l \in L$ and $|L| > 1$, where L is a set of disjoint labels: if $|L| = 2$ it is a *binary classification* problem, while if $|L| > 2$ it

is a *multi-class classification* problem. In multi-label scenarios, each example is associated with a set of labels Y , with $Y \subseteq L$ and $|L| > 1$, where L is a set of non-disjoint labels, and if $|L| = 2$ four cases can occur: the example does not belong to any class or belongs to both classes or belongs to one class or another hence two cases, but unlike binary classification, classes are not always mutually exclusive and the concepts of binary or multi-class classification are lost. A brief overview follows, while more comprehensive references are provided by Tsoumakas et al. (2010).

The multi-label classification problem is generally faced in literature with four distinct approaches known as *problem transformation* approaches. First attempts to approach multi-label problems consist in transforming problems in binary ones for each individual label, so any binary classifier can be applied, even different, depending on the label.

More sophisticated algorithms have tried to transform the multi-label problem into multiple multi-class or binary problems: typically such methods create multiple copies of the feature space or make several modifications to it in succession, exploiting ensembles in combination with SVMs, decision trees and probabilistic methods or boosting. But the literature identifies the inability to model the dependency between labels as the major limitation of these methods, also known as Binary Relevance (BR) methods.

To better model the dependence between labels by correlating different single-label classifier with each other, scientists developed the Chain Classifier method: the *chain* lies in the fact that classifier n will make its prediction based on both the inputs and the predictions of classifier $n - 1$.

Another approach, called LP, consists in transforming the multi-label problem into a single-class single-label problem, considering the powerset as a set of the 2^L possible combinations. The main advantage is to be able to directly model the dependencies between labels, obtaining better performance than BR, with the disadvantage of being easily applicable only when the computational complexity does not start to be particularly high making it necessary to think about how to deal with it. This work falls into this category and, for $|L| = 2$, provides a powerset $P = [0, 0], [0, 1], [1, 0], [1, 1]$ of size 4, where the element $[j, k]$ indicates whether the polarity is positive through j and whether the review is deceptive through k .

A variant of this method, called PairWise Transformation, exploits binary models trained for each pair so the predictions turn out to be sets

of pairs, but the approach can be extended to multi-label prediction, provided it takes into account the quadratic complexity involved.

While the illustrated approaches attempt to transform the problem, there are so-called *algorithm adaptation* approaches that try to directly exploit single-label methods in multi-label classification, such as Multi-Label k-Nearest Neighbors or the derived Instance-Based Learning by LR that also exploits LR, or Back-Propagation for Multi-Label Learning (BP-MLL). Instead, the criticality of representing the feature space by employing traditional machine learning methods in multi-label domains can now be easily circumvented managing the feature space through deep learning-based methods like BERT as proposed in this work.

2.3. Neural language models

Bengio et al. (2003) introduced the use of neural networks to build language models, i.e. probabilistic classifiers that learn to predict a probability distribution over sequences of words in order to represent the analyzed text, and outperformed previous statistical language models in the various branches of NLP, as extensively documented by Mulder et al. (2015).

The most important successive development in the field of language models has been the acquisition of the ability to analyze context: the processing of probability distributions takes into account not only the words that follow but also those that precede them, and in this way it is possible to better interpret the meaning of polysemous words and the syntax of texts, such as ELMo by Peters et al. (2018) or Flair by Akbik et al. (2018). In particular, many things changed when (Devlin et al., 2019) introduced the BERT language model, based on a neural architecture consisting of attention mechanisms (Vaswani et al., 2017), which shown to be able to match or exceed many of the state-of-the-art methods. The main advantage of BERT and its derivatives, in addition to the boost in contextuality capable of accounting for at least 512 tokens in the smaller versions of the model, lied in the possibility of pre-training the model (an operation that requires considerable computing resources and time) and then fine-tuning it for the specific task of interest: the latter process superficially, almost imperceptibly, modifies the pre-trained sub-word embeddings that nest in the BERT model, which can therefore be considered rather monolithic.

For these reasons, the approach proposed in this paper tries to make the most of the pre-trained model and the additional sentiment information by modifying the fine-tuning process as appropriate: acting directly on the pre-training phase to take sentiment into account would be tantamount to wasting the generalization effort made upstream in the construction of BERT. To date, to the best of our knowledge and although the scientific literature has largely emphasized the importance of sentiment for correctly detecting deception within reviews (Harris, 2012; Li et al., 2014; Ott et al., 2013), there is no scientific work in the literature that has attempted to combine sentiment classification capabilities in order to improve detection of deceptive reviews using Google BERT as proposed here: a work that could seem similar is the one of Shang et al. (2021) but in this case the sentiment information is built through an external model that has further labeled the YelpNYC dataset and modified the embeddings at the base of BERT incorporating the sentiment information, moreover it is not provided a precise information about the subdivision between training and testing used for the dataset, making difficult a direct comparison. In addition, differently from Ott et al. (2013, 2011), whose focus was mainly on developing separate classification models for positive and negative reviews, the proposed model exploits the contrast between positive and negative sentiment within a multi-label LP-type classification system in which such information is mixed with deceptiveness information so as to allow the underlying deep neural network to best extract the hidden features that correlate sentiment and deceptiveness for the purpose of revealing deception. In particular, Ott et al. (2013) claim to obtain comparable performance (equal or slightly lower based on the numbers provided)

by training a single classifier by merging datasets with positive and negative polarity: although not directly comparable in terms of technologies (shallow networks of NB and SVM type versus deep networks underlying BERT) and how the dataset is used (stratified 5 fold cross validation versus a more robust stratified 5 fold cross validation over 5 runs per 5 seeds), our proposal has instead been shown to benefit from combining such information.

3. Material and methods

This section introduces the materials and methods employed for the proposed study. Section 3.1 provides an overview of the datasets employed while Section 3.2 delves into aspects and features of the BERT language model and variants that were compared.

3.1. Dataset

In the following, the DOSC and YelpNYC datasets are illustrated in the 3.1.1 and 3.1.2 sections, respectively.

3.1.1. DOSC dataset

The *Deceptive Opinion Spam Corpus v1.4*² is constituted by truthful and deceptive reviews of 20 Chicago hotels. It was released, according to the sentiment of the reviews, in two papers: in particular positive sentiment reviews were discussed in Ott et al. (2011), while negative sentiment reviews in Ott et al. (2013).

The authors tried to maintain consistent data pre-processing procedures across the data: details about the differences were explained in the associated papers. To avoid vanishing this aspect of the original work, the only pre-processing applied here is related to enabling or not the lower-casing of the text, without removing any other character if not specified.

This corpus contains the following:

- 400 truthful positive reviews from TripAdvisor, described in Ott et al. (2011), mined following these criteria: 5-star reviews,³ written in English with at least 150 characters, from non first-time authors⁴;
- 400 deceptive positive reviews from Amazon Mechanical Turk⁵ (AMT) crowd-sourcing service, described in Ott et al. (2011);
- 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp, described in Ott et al. (2013), mined following these criteria: 1- or 2-star reviews;
- 400 deceptive negative reviews from AMT crowd-sourcing service, described in Ott et al. (2013).

Specifically, the above datasets (each consisting of 20 reviews for each of the 20 most popular Chicago hotels, the same for each dataset) were merged to create one omnicomprehensive, perfectly balanced, dataset of 1600 reviews. Hence, the labels considered were referred to deceptive or truthful and negative or positive.

Concerning the deceptive part of the dataset it is important to underline an aspect: in order to make it credible, the procedure to obtain such opinions followed the real one, asking those real authors of deceptive opinions to repeat their works again. In detail, the authors of the dataset created, altogether, 800 Human Intelligence Tasks (HITS) relying on the aforementioned AMT platform,⁶ proposing a \$1 reward for Turks residing in the US, with two restrictions: (1) 30 min to

² <https://myleott.com/op-spam.html>.

³ The score assigned by the user (from 1 star to 5 stars) shows the user's liking for the hotel being evaluated.

⁴ Generally, first-time authors are more likely to write deceptive reviews, the result of so-called shilling attacks (Wu et al., 2010).

⁵ <https://www.mturk.com/>.

⁶ 40 HITS per hotel, 20 positive and 20 negative.

Table 2
YelpNYC dataset: deceptiveness VS. sentiment polarity.

Sentiment polarity	Deceptiveness	
	Truthful	Deceptive
Positive	293,126	30,927
Negative	29,041	5958

complete the task and (2) a single author per review, to avoid confusing classifiers with different writing styles. Provided with the name and website of a hotel, Turks had to assume they worked for the listed hotel and write a deceptive review that was positive or negative and felt real. Finally, by filtering out the jobs that did not meet the requirements, e.g. because reviews were too short or copied, the 800 useful reviews were obtained.

3.1.2. YelpNYC dataset

The YelpNYC dataset, made available through the work of Rayana and Akoglu (2015), contains reviews of restaurants located in New York City. Specifically, the Yelp platform relies on a proprietary filtering algorithm that detects false and/or suspicious reviews and, although it makes them public, these are not placed on the Yelp “recommended reviews” page of the restaurant to which they refer, but instead they are to be retrieved through a specific link at the bottom of the page. Overall, this dataset contains a total of 359,052 reviews, both recommended and filtered (just over 10% of the total), which are respectively used as genuine and fake. In addition to reviews and their type (recommended or filtered), there are also information such as the username of the author, date and rating: the latter goes from 1 to 5 and it was decided to consider those with rating 1–2 and 3–5 respectively negative or positive from the point of view of sentiment polarity. The detail of the reviews contained in the dataset but obtained as a result of partitioning by deceptiveness and sentiment polarity is shown in Table 2.

3.2. BERT language model

The main architecture exploited in this article is the one created by Devlin et al. (2019), namely the Bidirectional Encoder Representations from Transformers, commonly abbreviated as BERT. As already mentioned, this language model exploits a technique of pre-training on a large corpus of text which allows it to acquire a high capacity of generalization. Subsequently, exploiting a technique defined as fine-tuning, it is possible to specialize the architecture employed for different processes of NLP on the basis of the needs, such as the Named Entity Recognition or the Relation Extraction, or even the Question Answering, the Sentiment Analysis or the Summarization: in particular, the technique of fine-tuning acts mainly on the last layers of the constituent neural network which are specialized for the specific task, with only slight modifications to the more internal layers which offer the greatest ability of generalization.

Main details about BERT architecture are given in Table 3. In detail, the transformer blocks, i.e. the hidden layers that constitute the transformer encoder, are 12 and 6 respectively, while the number of attention heads, often called self-attention (Vaswani et al., 2017), is 12 for both. Moreover, feed-forward networks hidden size is 768, while maximum sequence length parameter manageable by the models is 512: this essentially corresponds to the maximum acceptable size for the input vector. Consequently, the number of net weights is also significantly different due to the different number of hidden layers: 110 versus 66 million (M). The main hyper-parameters used for fine-tuning are always given in Table 3.

In order to properly handle sequences, BERT employs two particular tokens namely [SEP] and [CLS], which serve respectively to separate sentences and to provide an output vector of size equal to the hidden size H to be used as input for an arbitrary classifier, representing the entire sequence. The tokenization is done through a method called

Table 3
BERT_{Base} hyper-parameters.

Hyper-parameters	BERT _{Base}
Attention heads	12
Batch size	8
Epochs	5
Hidden size	768
Hidden layers	12
Learning rate	0.00001
Maximum sequence length	512
Parameters	110 M

WordPiece that aims to balance the problems of OOVs and vocabulary size by dividing words into sub-words common to many words (Wu et al., 2016).

Finally the final hidden state given by the first token used as input, i.e. the output of the transformers, it is indicated as a vector $C \in \mathbb{R}^H$. This vector is then inputted into the final fully-connected classification layer. Here, given $W \in \mathbb{R}^{K \times H}$ as parameter matrix of the classification layer, with K equal to the number of categories, the probability P for each category is calculated as:

$$P = \text{softmax}(CW^T) \quad (1)$$

The BERT model, as provided by the Hugging Face library⁷ employed here, involves the out-of-the-box use of the *Categorical Cross Entropy* loss function that is valid for multi-class classification. For the purpose of multi-label of LP-type classification it has been necessary to operate in a different way. In detail, avoiding to provide the model with the parameter related to the labels on which to act, it was possible to obtain in output only the logits, then properly manipulated for the multi-label classification using the loss function Binary Cross Entropy with Logits (BCEwL) provided by the torch library.⁸

In particular, the BCEwL loss function combines in one single class both a sigmoid layer and the Binary Cross Entropy loss function: their combination is more numerically stable than using them separately because combining the operations into one layer it is possible to take advantage of the LogSumExp function,⁹ improving numerical stability. The generic BCEwL can be described as:

$$l(x, y) = L = \{l_1, \dots, l_N\}^T, \quad (2)$$

$$l_n = -w_n [y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))]$$

where N is the batch size and i th targets should be numbers between 0 and 1.

For multi-label classification, BCEwL can be instead described as:

$$l_c(x, y) = L_c = \{l_{1,c}, \dots, l_{N,c}\}^T, \quad (3)$$

$$l_{n,c} = -w_{n,c} [p_c y_{n,c} \cdot \log \sigma(x_{n,c}) + (1 - y_{n,c}) \cdot \log(1 - \sigma(x_{n,c}))]$$

where c represents the class number ($c = 1$ for single-label binary classification, $c > 1$ for multi-label binary classification), n is the number of the sample in the batch and p_c is the weight of the positive answer for the class c : it makes possible to trade off precision and recall because $p_c < 1$ increases the precision, while $p_c > 1$ increases the recall. This last trick it is particularly important for unbalanced dataset and could be an alternative to stratification used for this work: for example, given a dataset with 400 positive and 800 negative examples of a single class, setting $p_c = 2$ would act as if the dataset contains 800 ($p_c \cdot$ positive examples) positive examples.

A brief overview of the proposed BERT-based fine-tuning process is given by Fig. 1. In the pre-processing phase, at the input of the labels on deceptiveness and sentiment polarity, a forced link is induced

⁷ https://huggingface.co/transformers/model_doc/bert.html.

⁸ <https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html>.

⁹ <https://en.wikipedia.org/wiki/LogSumExp>.

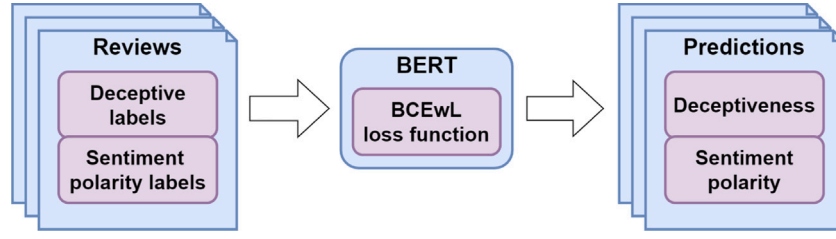


Fig. 1. Proposed BERT-based fine-tuning process.

through the use of the powerset in such a way that there is no longer a clear distinction between the options “Deceptiveness” and “Polarity”, but four options representing the relative possible combinations: this forces the deep neural network constituting BERT to give weight to both aspects. Therefore, through the BCEwL, it is trained a system in which the decrease of the loss is not tied only to one of the two labels but to their combination: the same output of the system is represented from a binary tuple that at the same time (1) maintains firm the link among the combinations being four the possible binary tuples in output and (2) allows the reconstruction of the single result related to the deceptiveness or to the polarity thanks to its nature of binary tuple.

Transformer. The transformer, introduced by Vaswani et al. (2017) is the most important component of the BERT architecture. If sequences of sub-words \mathbf{x} and \mathbf{y} are considered, the BERT architecture put the [CLS] token before \mathbf{x} and [SEP] after both \mathbf{x} and \mathbf{y} . E and LN will be the embedding function and the normalization layer respectively. The embedding is obtained as:

$$\hat{h}_i^0 = E(x_i) + E(i) + E(1_x) \quad (4)$$

$$\hat{h}_{j+|x|}^0 = E(y_j) + E(j + |x|) + E(1_y) \quad (5)$$

$$\hat{h}_i^0 = Dropout(LN(\hat{h}_i^0)) \quad (6)$$

Therefore the embedding encounters M transformer blocks where, said FF the Feed Forward layer, $GELU$ the element-wise Gaussian Error Linear Units activation function (Hendrycks & Gimpel, 2016) and $MHSA$ the Multi-Heads Self-Attention function respectively, for each of them it holds:

$$\hat{h}_i^{i+1} = Skip(FF, Skip(MHSA, \hat{h}_i^i)) \quad (7)$$

$$Skip(f, h) = LN(h + Dropout(f(h))) \quad (8)$$

$$FF(h) = GELU(h\mathbf{W}_1^\top + \mathbf{b}_1)\mathbf{W}_2^\top + \mathbf{b}_2 \quad (9)$$

where $h^i \in \mathbb{R}^{(|x|+|y|) \times d_h}$, $\mathbf{W}_1 \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_1 \in \mathbb{R}^{d_h}$, $\mathbf{W}_2 \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_2 \in \mathbb{R}^{d_h}$ and the new \hat{h}_i position is:

$$\begin{aligned} [\dots, \hat{h}_i, \dots] &= MHSA([h_1, \dots, h_{|x|+|y|}]) \\ &= \mathbf{W}_o \text{Concat}(h_1^1, \dots, h_i^N) + \mathbf{b}_o \end{aligned} \quad (10)$$

In each attention head it happens:

$$h_i^j = \sum_{k=1}^{|x|+|y|} Dropout(a_k^{(i,j)}) \mathbf{W}_V^j h_k \quad (11)$$

$$a_k^{(i,j)} = \frac{\exp\left(\frac{(\mathbf{W}_Q^j h_i)^T \mathbf{W}_K^j h_k}{\sqrt{d_h/N}}\right)}{\sum_{k'=1}^{|x|+|y|} \exp\left(\frac{(\mathbf{W}_Q^j h_i)^T \mathbf{W}_K^j h_{k'}}{\sqrt{d_h/N}}\right)} \quad (12)$$

where, said N the number of attention heads, $h_i^j \in \mathbb{R}^{(d_h/N)}$, $\mathbf{W}_o \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_o \in \mathbb{R}^{d_h}$ and $\mathbf{W}_Q^j, \mathbf{W}_K^j, \mathbf{W}_V^j \in \mathbb{R}^{d_h/N \times d_h}$.

4. Experiments and discussion

This section describes the experiments conducted to test the proposed approach. In particular, the experimental setup is outlined in Section 4.1, the evaluation metrics are described in Section 4.2 and, finally, the results achieved are discussed in Section 4.3. A snapshot of the inferential moment based on the proposed BERT-based fine-tuning process seen in the previous section is given by Fig. 2. In detail, a review whose deceptiveness is desired comes as input to the previously fine-tuned model. Then, the proposed approach makes sure that the information about deceptiveness and polarity of the review sentiment are linked, so the model provides a two-dimensional prediction: at this point the extraction of the required prediction, in this case the deceptiveness one, is performed.

4.1. Experimental setup

As said, the use of transformers-based language models such as BERT has become commonplace in the NLP domain, with simple fine-tuning of a pre-trained model often achieving performance equal to or better than the state of the art. The main problem in this practice lies in the instability of the process: the literature has pointed to several reasons behind this phenomenon, such as catastrophic forgetting issues and small size of fine-tuning datasets (Devlin et al., 2019; Dodge et al., 2020; Lee et al., 2020), or the use of a non-standard optimization method with biased gradient estimation, the limited reuse of important network chunks for downstream tasks, and the pre-determined use of few epochs for fine-tuning (Zhang et al., 2021). With the most recent studies (Mosbach et al., 2021; Zhang et al., 2021) it was possible to highlight how the instability of the process of fine-tuning is mainly due to the difficulties of optimization that lead to the vanishing of the gradient and the use, not always optimal, of a few epochs of fine-tuning, especially when the models used are large versions and the datasets used for fine-tuning are relatively small. For these reasons, since the case histories examined in this study are inherent, it was decided to extend the fine-tuning of the transformers-based language models analyzed to a number of epochs higher than the recommended 2–4 (Devlin et al., 2019), going to expand the literature with a further case study specific to the task of recognition of fake reviews: in detail 5, 10, 15 and 20 epochs were experimented.

Regarding the DOSC dataset, the experiments were performed by breaking the original dataset into train and validation datasets according to the percentages generally employed in the literature: the training set is progressively increased from 50% to 90% with a step of 10 percentage points (consequently the validation set is accordingly reduced from 50% to 10% of the unique original dataset). Moreover, for each model, 5 different datasets have been created exploiting random seeds from 0 to 4 and employing the stratification (Sechidis et al., 2011) to maintain the balance of the dataset between training and validation datasets. Therefore, for each dataset 5 runs per model were performed, obtaining 25 experiments each of which performed in cross validation with $k = 5$ in order to mitigate overfitting effects due to the limited size of the dataset used. In addition, an extra dropout layer with $p = 0.1$ was added before the terminal classification layer.

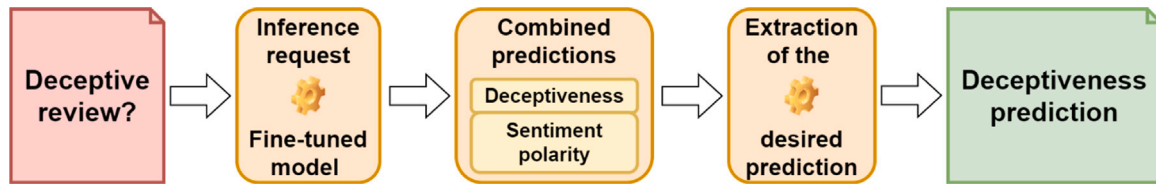


Fig. 2. Snapshot of the inferential moment.

Regarding the YelpNYC dataset, the experiments were carried out in an analogous way to the DOSC dataset, with the only exception related to the subdivision between training and validation, chosen to be equal to 90/10. In addition, in view of the strong imbalance as evident from Table 2, the dataset was balanced against the least numerous class so that there were 5958 reviews for every possible combination of deceptiveness and sentiment polarity, for a total of 23832 reviews after its pre-processing: given the order of magnitude more than the DOSC dataset, this is still an additional benchmark to test and validate the proposed approach with a larger dataset.

4.2. Evaluation metrics

Due to the different nature of multi-label classification problems, the metrics for evaluating label classification performance are different. In general, given T and P the sets of true and predicted labels for a given example, it can be identified two main metrics:

- Precision, Recall and F_1 score, defined respectively as $P \triangleq \frac{|T \cap P|}{|P|}$, $R \triangleq \frac{|T \cap P|}{|T|}$ and their harmonic mean $F_1 \triangleq \frac{2 \cdot P \cdot R}{P + R}$.
- Exact match (also known as subset accuracy or simply accuracy) indicates the percentage of correctly classified labels sets, hence it is the strictest metric.

In the case of multi-class or multi-label problems, it is necessary to define how to proceed to the operation of average among the different samples. The most diffused methodologies are:

- Binary. It reports results for the specified class but it is applicable only if target labels are binary.
- Micro. It calculates metrics globally by counting the total true positive, false both positives and negatives.
- Macro. It calculates metrics for each label finding it unweighted mean: it should be avoided for imbalanced datasets.
- Weighted (Macro). It calculates metrics for each label finding their average weighted by the number of true instances for each label, taking into account a possible unbalance of the dataset.

In the specific case of the employed dataset, the most popular metrics in the literature are Accuracy, i.e. exact match, and F_1 score generally used with binary average. In addition to these, Micro F_1 will also be used to report the result of the multi-label of LP-type scenario, i.e. the labels *deceptive* and *polarity* combined.

4.3. Results and discussion

Starting from the DOSC dataset, F_1 and Accuracy reported in Table 4 and, for the sake of completeness, related Precision, Recall and TNR are reported in Table 5. These results show that generally there is a better performance with a training set between 80% and 90%: as expected, in scenarios with small datasets, the increase of data in the training phase impacts the performance of the model, finding a behavior similar to the literature trend observed in Tables 6 and 7, adding our results to that reported by Fahfouh et al. (2020).

Going into the detail of the comparison between the single label methodology, which in this case refers to the deceptiveness or not of the review, and the multi-label one, it is good to clarify the results: in

particular it is the column with subscript d that should be compared with the single label in order to understand if setting the classification in a multi-label way has better effects on the performance of the model. In fact the results seem to confirm this hypothesis, being the performances (also not considering the best case, but comparing the results with the same Train Size) almost always higher. The BERT cased model, that is able to distinguish between lower and upper case, obtains higher performances than the uncased and better in absolute. In addition, it is interesting to note an aspect: a reduction in performance in the recognition of deceptiveness does not always and necessarily correspond to a lower ability to recognize the polarity of the sentiment and, in order to further investigate this case study, it would be necessary to have datasets for deceptive reviews able to exploit multitask learning capacities due to the existence of several similar tasks (e.g. emotion detection and sentiment analysis or offensive language identification or hate speech detection) as it has been already proposed in literature (Barbieri et al., 2020) hence confirm a possible correlation between emotions and deceptive reviews.

Comparison with the state-of-the-art. To better understand how much the proposed method could improve on the previous ones, Tables 6 and 7 show the results of the current states-of-the-art on the dataset under study (Cagnina & Rosso, 2015; Fahfouh et al., 2020; Feng, Banerjee et al., 2012; Ren & Ji, 2017; Zhang et al., 2018). In detail, Fahfouh et al. (2020) proposed a model based on Paragraph Vector Denoising Autoencoder (PV-DAE) with the goal of capturing not only the semantic aspect but also the context of each opinion and the results obtained underlined the importance of the consideration at the base of their work. Instead, Zhang et al. (2018) proposed an approach called Deceptive Review Identification by Recurrent Convolutional Neural Network (DRI-RCNN) to identify deceptive reviews using deep learning and word contexts. Ren and Ji (2017) exploiting a Gated Recurrent Neural Network (GRNN) based model capable of modeling discourse information and producing a vector of documents, again for the purpose of identifying deceptive reviews. Finally, the works of Cagnina and Rosso (2015) and Feng, Banerjee et al. (2012), based on Stylistic (n-gram), Emotional and Linguistic features (SEL-feature) and Syntactic Stylometry for Deception Detection (SSDD) features respectively, constitute a baseline often used in literature, already discussed in Section 2.1.1.

Analyzing the different models, it is clear that the size of the training datasets directly affects performance, both in terms of accuracy and F_1 as shown in Tables 6 and 7. All models improve both metrics from 50% to 60%. At 70% the F_1 of the SEL-feature model is the only one to record a decline that will increase with the size of the training dataset, and similarly for the accuracy from 80%. In contrast, the SSDD model sees a decrease in F_1 starting at 80% and, as with the SEL-feature model, the accuracy only begins to decrease as the training set size increases by an additional 10%, i.e. at 90%. The GRNN, DRI-RCNN, and PV-DAE models always improve performance in both metrics as the size of the training dataset increases. For the same size, PV-DAE is better than DRI-RCNN and the latter is better than GRNN. Finally, regarding the best proposed model, it is interesting to note the following: the trend for both metrics with a training dataset size of 60%, 70% and 80% is oscillating around average values of accuracy and F_1 equal to 0.9274 and 0.9299, respectively. Only by increasing the size of the training dataset to 90% is it possible to obtain an increase of about 1 percentage point, with values of accuracy and F_1 equal to

Table 4

Best results on DOSC dataset. Superscripts B, M indicate whether the metric is Binary or Micro, respectively. In contrast, subscripts m, d, p indicate whether the result is relative to the multi-label or to the deceptive or polarity labels extracted from the multi-label experiment, respectively.

Model	Train size	Single label				Multi label					
		F_{ld}^B	Acc_d^B	F_{lp}^B	Acc_p^B	F_{lm}^B	Acc_m^B	F_{ld}^B	Acc_d^B	F_{lp}^B	Acc_p^B
BERT ^{Cased} _{Base}	0.5	86.50	85.38	97.68	97.36	92.67	85.50	87.80	87.88	97.42	97.38
	0.6	86.96	85.99	98.35	98.20	95.50	91.25	92.92	92.81	98.12	98.13
	0.7	88.36	87.64	98.78	98.51	94.72	89.58	93.12	92.92	96.33	96.25
	0.8	88.42	87.55	99.01	98.76	95.74	91.56	92.94	92.50	98.74	98.75
	0.9	87.50	86.10	99.12	98.99	96.62	93.13	93.98	93.75	99.37	99.38
BERT ^{Uncased} _{Base}	0.5	87.02	85.99	97.55	97.42	93.81	87.50	90.25	89.63	97.54	97.50
	0.6	88.09	87.38	98.02	97.88	93.50	86.72	89.71	89.06	97.52	97.50
	0.7	86.69	85.07	98.18	98.01	92.32	84.17	88.09	86.88	96.92	96.88
	0.8	88.36	87.39	98.32	98.19	92.88	85.31	88.20	86.88	98.11	98.13
	0.9	88.83	87.90	98.56	98.33	94.30	88.75	91.25	91.25	97.44	97.50

Table 5

Precision (P) and Recall (R) related to the best results obtained. Superscripts B, M indicate whether the metric is Binary or Micro, respectively. In contrast, subscripts m, d, p indicate whether the result is relative to the multi-label or to the deceptive or polarity labels extracted from the multi-label experiment, respectively.

Model	Train size	Single label						Multi label								
		P_d^B	R_d^B	TNR_d^B	P_p^B	R_p^B	TNR_p^B	P_m^B	R_m^B	TNR_m^B	P_d^B	R_d^B	TNR_d^B	P_p^B	R_p^B	TNR_p^B
BERT ^{Cased} _{Base}	0.5	86.32	86.68	88.73	98.08	97.28	97.58	93.21	92.14	90.00	86.85	88.77	87.60	97.19	97.65	97.48
	0.6	87.86	86.07	86.39	97.33	99.39	99.58	96.61	94.41	96.57	92.34	93.51	92.33	96.94	99.33	97.47
	0.7	89.28	87.46	88.71	99.43	98.14	99.15	93.49	95.98	98.06	93.48	92.76	92.27	95.61	97.06	97.59
	0.8	89.52	87.34	87.90	98.79	99.23	99.35	96.65	94.85	92.65	91.92	93.99	95.41	98.47	99.01	97.28
	0.9	87.24	87.77	86.06	99.85	98.40	96.98	97.37	95.88	95.22	94.92	93.06	93.71	98.82	99.93	99.08
BERT ^{Uncased} _{Base}	0.5	87.21	86.83	84.77	98.14	96.97	97.64	93.06	94.57	92.09	89.17	91.36	91.64	97.37	97.71	98.04
	0.6	87.49	88.70	91.15	98.24	97.80	97.88	93.95	93.05	93.55	90.17	89.25	89.53	96.99	98.06	96.90
	0.7	87.42	85.97	86.63	99.09	97.28	94.93	92.67	91.97	89.54	88.29	87.89	88.85	97.75	96.10	96.08
	0.8	87.24	89.51	88.60	97.95	98.69	99.90	93.03	92.73	91.50	87.26	89.16	88.31	98.61	97.61	95.98
	0.9	88.79	88.87	86.71	98.32	98.88	99.92	93.13	95.50	95.22	91.99	90.52	90.12	98.66	96.25	96.86

Table 6

Accuracy states-of-the-art.

Model	50%	60%	70%	80%	90%
Proposed	0.8788	0.9281	0.9292	0.9250	0.9375
PV-DAE (Fahfouh et al., 2020)	0.8575	0.8750	0.8833	0.8875	0.9250
DRI-RCNN (Zhang et al., 2018)	0.8293	0.8415	0.8589	0.8724	0.8815
GRNN (Ren & Ji, 2017)	0.8015	0.8189	0.8324	0.8415	0.8582
SEL-feature (Cagnina & Rosso, 2015)	0.8353	0.8481	0.8510	0.8434	0.8314
SSDD (Feng, Banerjee et al., 2012)	0.8487	0.8512	0.8558	0.8609	0.8571





Table 7

F_{ld}^B states-of-the-art.

Model	50%	60%	70%	80%	90%
Proposed	0.8780	0.9292	0.9312	0.9294	0.9398
PV-DAE (Fahfouh et al., 2020)	0.8492	0.8742	0.8828	0.8860	0.9240
DRI-RCNN (Zhang et al., 2018)	0.8123	0.8301	0.8458	0.8536	0.8659
GRNN (Ren & Ji, 2017)	0.8037	0.8281	0.8332	0.8417	0.8513
SEL-feature (Cagnina & Rosso, 2015)	0.8203	0.8492	0.8432	0.8321	0.8230
SSDD (Feng, Banerjee et al., 2012)	0.8203	0.8311	0.8438	0.8312	0.8210

Table 8

Number of incorrectly predicted reviews.

Typology	Unidentified deceptiveness	Unidentified polarity
 Deceptive negative	8	0
 Deceptive positive	5	2
 Truthful negative	7	3
 Truthful positive	10	2

0.9375 and 0.9398 respectively. State of the art models for deceptive opinion spam recognition are based on traditional machine learning or deep learning that exploits language models and word embeddings for text representation, capturing context, semantics, polysemy and morphosyntactic variations. These models face several challenges as shown in the introduction and related work, but more importantly, employing large models or models with even more parameters available is not always worth it compared to the performance gains possible. Rather, as

this work shows, it is possible to adopt a model among those already widely used and to change its scope, trying to exploit the correlation between the deceptiveness of an opinion and its various other aspects, including psychological ones such as emotion and its polarity. In this way it is possible to obtain an increase in performance with practically the same amount of resources employed, bypassing the constraint on performance and exploiting additional contextual information otherwise left out: by exploiting the superior context analysis capability of BERT, which in the basic models extends up to 512 parallel processable tokens through appropriate setting of the maximum sequence length hyper-parameter, it is possible to extract hidden features not only automatically due to the depth of the model but also by having the ability to fully analyze reviews without the limitations of a reduced context window, as is the case in Fahfouh et al. (2020).

This explains the state-of-the-art performance in terms of accuracy and F_1 achieved by the proposed model.

Table 9

Reviews with the worst predictions among the most challenging.

Text	D	P
I work for a software marketing firm, and that job requires me to travel at least 100 days per year. I spend a large amount of time in hotels. I normally don't write reviews for hotels, but the Homewood Suites by Hilton were above and beyond. Let's start where it matters: the rooms. They are top notch. There's nothing better than walking in to a clean, beautiful rooms after a long day of travel. The beds are among the best I've ever slept in hotels. The rest of the room is just as nice and makes me feel at home. The price: Low. I looked for similar rooms in the area at other hotels, and everything else was at least 10% more expensive. But you don't get what you pay for here. You get much more than that! The location: Couldn't be better. You want to go to a world class restaurant? That's within walking distance. You want a great shopping experience? Again, walking distance. Anything else you want? Chances are, it's within walking distance. Overall, one of the best hotels I've ever stayed in.	0	4
This hotel is the perfect location for downtown Chicago shopping. The only thing is the pool is extremely small — it is indoors, but looks much larger on the website.	3	0
At the InterContinental Chicago, you pay for more than a room. You pay for an experience. Let me elaborate. I was a bit reluctant to pay a premium for location and amenities. Like any novice traveller to Chicago, I saw the mass glut of cheap accommodations out in the suburbs, and those were originally what I set my sights on. Price is a powerful signal, and the sheer discount of those options drew me in. But after doing a bit of research, I realised that distance does have its costs. The price of transit in the Chicagoland area isn't free. And it's far from what I'm used to. Car use is price prohibitive. (I knew that originally, which is why I wasn't going to drive.) But the sting of using the automobile (parking, congestion, toll roads, etc.) gets shoved off onto taxi fares and auto rental. So, I ultimately had to end up looking for a place well connected with mass transit. But mass transit has its costs too, which meant that staying at a focal point downtown had its advantages. That's what originally drew me to InterContinental. But InterContinental is much more than the right room at the right place. The decor of the rooms and the amenities available to the guest are unsurpassable. I was travelling on international business, so the ease with which I could transact in foreign currency was a draw. (They had a currency exchange right there on the premises.) And the fact that massages, dry cleaning, and event planning were available in-house was a huge plus. My work gets me tied up in knots. It also leaves little time for doing chores. So the ability to grab a massage while my suits were being finished was a relief for me. But, on top of all that, the event planning was what put InterContinental over the top. Being able to coordinate a comprehensive tour of Chicago while completing my business and managing my personal life was a weight off my shoulders- and something I won't forget. In the end, if you're a newcomer to Chicago staying in a time-crunch, the InterContinental is for you. The pricetag of rooming downtown can seem prohibitive, but when you factor in the convenience and savings in marginal costs, it's nothing in the end. And the InterContinental is not only price-competitive with the other downtown options, it also makes staying downtown the convenience and thrill it should be.	4	0
The room was very spacious with very nice, colorful decorating. The beds felt like lying on a cloud, and there were four pillows on each bed, two fiberfill and two down so you could choose which you like better. The bathroom was very nice also. Overall, I was very pleased with the stay. You had to show your room key to get in after ten p.m. I thought that was a good security issue. It has a great location too, close to downtown. I would recommend it. The decorations were very classy. Most hotels have two of the same framed art, boring bedspreads and lumpy pillows. The Hyatt Regency was just the opposite.	1	0

Challenging entities. Starting from the two labels used, i.e. *deceptiveness* and *sentiment polarity*, it is possible to obtain 4 combinations that are perfectly distributed within the original dataset (400 reviews for each): this fair balance has been preserved in each of the validation datasets, extrapolated by varying both the overall percentage of the original dataset dedicated to training and the five random seeds used for generation. Analyzing in detail the 5 validation datasets generated when the size of the training dataset corresponded to 90% of the original dataset, it is possible to observe the distribution of never correctly predicted items, both in terms of deceptiveness and sentiment polarity, expressed in Table 8. The column *Typology* shows the four possible combinations with the two labels used in the ground truth dataset and, in addition, the color next to each pair of labels will be used to identify in subsequent tables the true pair to which the specific review belongs. On the other hand, the column *Unidentified deceptiveness* indicates the number of sentences out of a total of 800 (160 for each of the 5 validation datasets) whose label “Deceptiveness” was never correctly predicted in the 5 attempts made for each random seed. The same applies to the *Unidentified polarity* column.

What is interesting to note is that, while the prediction is never correct for one label, there is an always correct prediction (5 out of 5) for the other label in almost all cases: the only exceptions are the four reviews shown in Table 9.

Firstly, there is the color indication introduced with the previous table, then the text of the review followed by the number of correct predictions out of 5 for both the label *deceptiveness* in column *D* and the label *sentiment polarity* in column *P*. It is possible to observe that these reviews are of different lengths, so any architectural limitations dictated by the need to truncate excessively long sentences are not so influential: probably the link between deceptiveness and sentiment,

when useful, is already evident as the review proceeds therefore not polarized at the beginning, in the middle or at the end of it. Furthermore, looking at the number of predictions, it would seem that while a failure to predict deceptiveness can still be attributed to a rather good sentiment prediction (4 out of 5 for the first review in Table 9), the same is not always true in the opposite case where fluctuating situations are obtained, with 3, 4 or only one correct prediction out of 5 (second, third and fourth review): the last is also the most interesting case. In fact, looking carefully at the structure of the review, that indicates an important option to try to improve performance: the review in question seems a positive review for any reader, until it reaches the final sentence *The Hyatt Regency was just the opposite*. which completely negates the previous text. This suggests that the reviewer wanted to somehow ridicule the subject of the review: this aspect is satirical in nature. Consequently, in addition to sentiment, the ability to identify satirical aspects could help in recognizing deceptive reviews: together with an ad-hoc integrated label taken into account in the calculation of deceptiveness, an analysis capable of separating parts of text of individual reviews, identifying the parts of the text subject to satire (in this case the whole review except the last sentence) and the parts of the text which introduce the phenomenon itself (in this case the last sentence) could be useful.

Insight into the YelpNYC dataset. Although the size of this dataset does not allow for a punctual qualitative analysis like the one just seen for the DOSC dataset, it is possible to dwell on a quantitative analysis, to which some considerations regarding comparison with the state of the art have been added. The results obtained are reported in Table 10 (F_1 and accuracy) and Table 11 (related precision, recall and TNR). BERT^{Cased}_{Base} shows an average improvement of 5.03% in F_1 score and 4.58% in accuracy with respect to deceptiveness detection, confirming

Table 10

Best results on YelpNYC dataset. Superscripts *B*, *M* indicate whether the metric is Binary or Micro, respectively. In contrast, subscripts *m*, *d*, *p* indicate whether the result is relative to the multi-label or to the deceptive or polarity labels extracted from the multi-label experiment, respectively.

Model	Single label				Multi label					
	F_{ld}^B	Acc_d^B	F_{lp}^B	Acc_p^B	F_{lm}^M	Acc_m^M	F_{ld}^B	Acc_d^B	F_{lp}^B	Acc_p^B
BERT ^{Cased} _{Base}	68.46	67.74	94.22	93.75	83.43	68.62	73.49	72.32	93.83	93.83
BERT ^{Uncased} _{Base}	68.32	67.57	93.98	93.61	83.38	68.51	72.32	72.13	93.76	93.78

Table 11

Precision (P) and Recall (R) related to the best results obtained. Superscripts *B*, *M* indicate whether the metric is Binary or Micro, respectively. In contrast, subscripts *m*, *d*, *p* indicate whether the result is relative to the multi-label or to the deceptive or polarity labels extracted from the multi-label experiment, respectively.

Model	Single label						Multi label								
	P_d^B	R_d^B	TNR_d^B	P_p^B	R_p^B	TNR_p^B	P_m^M	R_m^M	TNR_m^M	P_d^B	R_d^B	TNR_d^B	P_p^B	R_p^B	TNR_p^B
BERT ^{Cased} _{Base}	68.22	68.71	67.60	94.38	94.06	94.73	82.77	84.10	82.80	73.33	73.66	73.90	93.19	94.48	93.03
BERT ^{Uncased} _{Base}	68.60	68.05	68.97	93.50	94.46	95.31	83.59	83.17	83.98	73.49	71.19	73.37	92.91	94.63	97.11

the goodness of the proposed method within the procedure adopted for the use of the YelpNYC dataset. With similar performance, the trend is also confirmed by the uncased version.

Indeed, the weakness of the use of the YelpNYC dataset in the literature is largely related to the absence of an official division between training and validation data: this absence makes a direct comparison difficult with many of the more recent works (Hegde et al., 2021; Khalifa et al., 2021; Shang et al., 2021), where this information is missing or incomplete. Secondly, due to this lack, the imbalance between positive and negative reviews (the latter are about 10% of the total) often makes the presence of the accuracy result vain if not accompanied by the F_1 score: where such considerations have been made, the results are considered insufficient even in the face of possible expedients such as undersampling and oversampling techniques adopted by Budhi et al. (2021).

5. Conclusion

In this study, a new methodology was proposed to approach the recognition of deceptive opinions, starting from a dataset in which the polarity (positive or negative) of the sentiment of reviews, whether they are truthful or deceptive, is present. Specifically, this paper provides a multi-label model of LP-type for fake review detection able to exploit the expressive power of text representation of transformers-based language models such as BERT and derivatives by correlating, then taking into account during the classification process, also the sentiment: in fact, several studies have shown that sentiment and words with stronger polarities are often a sign of deceptiveness of a review. The results were particularly encouraging, showing how the link between sentiment and opinion truthfulness can be captured by multi-label approaches.

Furthermore, the proposed study should be analyzed in light of a number of limitations. First of all, the size of the dataset employed is modest, so there could be consequent distortions in the results, but, in this regard, it is also important to emphasize the inherent difficulty in collecting a manually annotated dataset in which there are both true and deceptive reviews, whatever the area studied. Secondly, the approach employed is centered on reviews, so if on the one hand the possibility of exploiting the information content due to the polarity of sentiment was explored, on the other hand it could be further improved by extending the dataset with information on user profiles and their networking activities as a further discriminator for the recognition of fakes. Moreover, the used dataset labels the sentiment in reviews exclusively as positive or negative instead more granularity of classification could help.

Thus, in the future, it might be interesting to consider multi-task learning situations (Barbieri et al., 2020) by extending the dataset and validating the proposed method in a more robust way (e.g., by intervening also on additional BERT hyper-parameters), perhaps employing other emotional states (e.g., happiness, sadness, anger, satire,

humor) and conducting a more granular analysis by distinguishing not only between positive and negative reviews but with different ratings (e.g., from 1 to 5 stars) so as to be able to employ other datasets and perhaps obtain more satisfactory performance overall. A further cue for future development concerns the possibility of a comparison between different models in single-label and multi-class classification modes, so as to understand if and how this contamination of information can improve the performance of different models. Finally, it may be interesting to test the effectiveness of algorithmic solutions such as BP-MLL (Zhang & Zhou, 2006) in a multi-task learning scenario.

CRedit authorship contribution statement

Rosario Catelli: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Hamido Fujita:** Writing – review & editing, Supervision. **Giuseppe De Pietro:** Resources, Writing – review & editing, Supervision, Funding acquisition. **Massimo Esposito:** Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., & Vigna, G. (2018). Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE security and privacy workshops, SP workshops 2018, San Francisco, CA, USA, May 24, 2018* (pp. 89–95). IEEE Computer Society, <http://dx.doi.org/10.1109/SPW.2018.00022>.
- Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018* (pp. 1638–1649). Association for Computational Linguistics, URL: <https://www.aclweb.org/anthology/C18-1139/>.
- Askalidis, G., & Malthouse, E. C. (2016). The value of online customer reviews. In S. Sen, W. Geyer, J. Freyne, & P. Castells (Eds.), *Proceedings of the 10th ACM conference on recommender systems, Boston, MA, USA, September 15–19, 2016* (pp. 155–158). ACM, <http://dx.doi.org/10.1145/2959100.2959181>.
- Banerjee, S., & Chua, A. Y. K. (2014). A theoretical framework to identify authentic online reviews. *Online Information Review*, 38(5), 634–649. <http://dx.doi.org/10.1108/OIR-02-2014-0047>.

- Banerjee, S., Chua, A. Y., & Kim, J. (2015). Using supervised learning to classify authentic and fake online reviews. In D. S. Kim, S. Kim, S. Lee, L. Hanzo, & R. Ismail (Eds.), *Proceedings of the 9th international conference on ubiquitous information management and communication, IMCOM 2015, Bali, Indonesia, January 08 - 10, 2015* (pp. 88:1–88:7). ACM, <http://dx.doi.org/10.1145/2701126.2701130>.
- Barbieri, F., Camacho-Collados, J., Anke, L. E., & Neves, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of ACL: vol. EMNLP 2020, Findings of the association for computational linguistics: EMNLP 2020, Online Event, 16-20 November 2020* (pp. 1644–1650). Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/2020.findings-emnlp.148>.
- Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155, URL: <http://jmlr.org/papers/v3/bengio03a.html>.
- Budhi, G. S., Chiong, R., & Wang, Z. (2021). Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications*, 80(9), 13079–13097. <http://dx.doi.org/10.1007/s11042-020-10299-5>.
- Cagnina, L. C., & Rosso, P. (2015). Classification of deceptive opinions using a low dimensionality representation. In A. Balahur, E. V. der Goot, P. Vossen, & A. Montoyo (Eds.), *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis, WASSA@EMNLP 2015, 17 September 2015, Lisbon, Portugal* (pp. 58–66). The Association for Computer Linguistics, <http://dx.doi.org/10.18653/v1/w15-2909>.
- Cai, L., Song, Y., Liu, T., & Zhang, K. (2020). A hybrid BERT model that incorporates label semantics via adjustable attention for multi-label text classification. *IEEE Access*, 8, 152183–152192. <http://dx.doi.org/10.1109/ACCESS.2020.3017382>.
- Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Najada, H. A. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2, 23. <http://dx.doi.org/10.1186/s40537-015-0029-9>.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (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)* (pp. 4171–4186). Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/n19-1423>.
- Ding, J., Hu, Y., & Chang, H. (2020). BERT-based mental model, a better fake news detector. In *ICCAI '20: 2020 6th international conference on computing and artificial intelligence, Tianjin, China, April 23-26, 2020* (pp. 396–400). ACM, <http://dx.doi.org/10.1145/3404555.3404607>.
- Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., & Smith, N. A. (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *CoRR abs/2002.06305*. URL: <https://arxiv.org/abs/2002.06305>, [arXiv:2002.06305](https://arxiv.org/abs/2002.06305).
- Fahfouh, A., Riffi, J., Mahraz, M. A., Yahyaouy, A., & Tairi, H. (2020). PV-DAE: A hybrid model for deceptive opinion spam based on neural network architectures. *Expert Systems with Applications*, 157, Article 113517. <http://dx.doi.org/10.1016/j.eswa.2020.113517>.
- Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting burstiness in reviews for review spammer detection. In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, & I. Soboroff (Eds.), *Proceedings of the seventh international conference on weblogs and social media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press, URL: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6069>.
- Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic stylometry for deception detection. In *The 50th annual meeting of the association for computational linguistics, proceedings of the conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short papers* (pp. 171–175). The Association for Computer Linguistics, URL: <https://www.aclweb.org/anthology/P12-2034/>.
- Feng, S., Xing, L., Gogar, A., & Choi, Y. (2012). Distributional footprints of deceptive product reviews. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, & Z. Tufekci (Eds.), *Proceedings of the sixth international conference on weblogs and social media, Dublin, Ireland, June 4-7, 2012*. The AAAI Press, URL: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4675>.
- Harris, C. G. (2012). Detecting deceptive opinion spam using human computation. In Y. Chen, P. G. Ipeirotis, E. Law, L. von Ahn, & H. Zhang (Eds.), *AAAI workshops: vol. WS-12-08, The 4th human computation workshop, HCOMP@AAAI 2012, Toronto, Ontario, Canada, July 23, 2012*. AAAI Press, URL: <http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/view/5256>.
- Hegde, S., Raj Rai, R., Sunitha Hiremath, P. G., & Gangisetty, S. (2021). Fake review detection using hybrid ensemble learning. In S. M. Thampi, E. Gelenbe, M. Atiquzzaman, V. Chaudhary, & K.-C. Li (Eds.), *Advances in computing and network communications* (pp. 259–269). Singapore: Springer Singapore, http://dx.doi.org/10.1007/978-981-33-6987-0_22.
- Hendrycks, D., & Gimpel, K. (2016). Bridging nonlinearities and stochastic regularizers with Gaussian error linear units. *CoRR abs/1606.08415*. URL: <http://arxiv.org/abs/1606.08415>, [arXiv:1606.08415](https://arxiv.org/abs/1606.08415).
- Hernández-Castañeda, A., Calvo, H., Gelbukh, A. F., & Flores, J. J. G. (2017). Cross-domain deception detection using support vector networks. *Soft Computing*, 21(3), 585–595. <http://dx.doi.org/10.1007/s00500-016-2409-2>.
- Hosseini, M., & Mukherjee, A. (2017). Detecting sockpuppets in deceptive opinion spam. In A. F. Gelbukh (Ed.), *Lecture notes in computer science: vol. 10762, Computational linguistics and intelligent text processing - 18th international conference, CICLING 2017, Budapest, Hungary, April 17-23, 2017, Revised selected papers, Part II* (pp. 255–272). Springer, http://dx.doi.org/10.1007/978-3-319-77116-8_19.
- Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In M. Najork, A. Z. Broder, & S. Chakrabarti (Eds.), *Proceedings of the international conference on web search and web data mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008* (pp. 219–230). ACM, <http://dx.doi.org/10.1145/1341531.1341560>.
- Jindal, N., Liu, B., & Lim, E. (2010). Finding unusual review patterns using unexpected rules. In J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, & A. An (Eds.), *Proceedings of the 19th ACM conference on information and knowledge management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010* (pp. 1549–1552). ACM, <http://dx.doi.org/10.1145/1871437.1871669>.
- Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8), 11765–11788. <http://dx.doi.org/10.1007/s11042-020-10183-2>.
- Kennedy, S., Walsh, N., Sloka, K., McCarren, A., & Foster, J. (2019). Fact or factitious? Contextualized opinion spam detection. In F. E. Alva-Manchego, E. Choi, & D. Khashabi (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student research workshop* (pp. 344–350). Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/p19-2048>.
- Khalifa, M. B., Elouedi, Z., & Lefèvre, E. (2021). Evidential spammers and group spammers detection. In A. Abraham, N. Gandhi, T. Hanne, T. Hong, T. N. Rios, & W. Ding (Eds.), *Lecture notes in networks and systems: vol. 418, Intelligent systems design and applications - 21st international conference on intelligent systems design and applications (ISDA 2021) Held during December 13-15, 2021* (pp. 255–265). Springer, http://dx.doi.org/10.1007/978-3-030-96308-8_23.
- Kim, J., Kang, J., Shin, S., & Myaeng, S.-H. (2021). Can you distinguish truthful from fake reviews? User analysis and assistance tool for fake review detection. In *Proceedings of the first workshop on bridging human-computer interaction and natural language processing* (pp. 53–59). Online: Association for Computational Linguistics, URL: <https://www.aclweb.org/anthology/2021.hcinlp-1.9>.
- Kumar, N., Venugopal, D., Qiu, L., & Kumar, S. (2018). Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems*, 35(1), 350–380, URL: <http://www.jmis-web.org/articles/1370>.
- Lee, C., Cho, K., & Kang, W. (2020). Mixout: Effective regularization to fine-tune large-scale pretrained language models. In *8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, URL: <https://openreview.net/forum?id=HkgaETNdB>.
- Lee, K., Webb, S., & Ge, H. (2014). The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdurfing. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, & A. H. Oh (Eds.), *Proceedings of the eighth international conference on weblogs and social media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press, URL: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8078>.
- Li, J., Cardie, C., & Li, S. (2013). TopicSpam: a topic-model based approach for spam detection. In *Proceedings of the 51st annual meeting of the association for computational linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short papers* (pp. 217–221). The Association for Computer Linguistics, URL: <https://www.aclweb.org/anthology/P13-2039/>.
- Li, F., Huang, M., Yang, Y., & Zhu, X. (2011). Learning to identify review spam. In T. Walsh (Ed.), *IJCAI 2011, Proceedings of the 22nd international joint conference on artificial intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011* (pp. 2488–2493). IJCAI/AAAI, <http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-414>.
- Li, J., Ott, M., Cardie, C., & Hovy, E. H. (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd annual meeting of the association for computational linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long papers* (pp. 1566–1576). The Association for Computer Linguistics, <http://dx.doi.org/10.3115/v1/p14-1147>.
- Martínez-Torres, M., & Toral, S. (2019). A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management*, 75, 393–403. <http://dx.doi.org/10.1016/j.tourman.2019.06.003>.
- Maslowska, E., Malthouse, E. C., & Viswanathan, V. (2017). Do customer reviews drive purchase decisions? The moderating roles of review exposure and price. *Decision Support Systems*, 98, 1–9. <http://dx.doi.org/10.1016/j.dss.2017.03.010>.
- Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jarrarweh, Y., & Maqsood, S. (2021). Fake reviews detection: A survey. *IEEE Access*, 9, 65771–65802. <http://dx.doi.org/10.1109/ACCESS.2021.3075573>.
- Molla, A., Biadgie, Y., & Sohn, K.-A. (2017). Detecting negative deceptive opinion from tweets. In *Mobile and wireless technologies 2017* (pp. 329–339). Springer Singapore, http://dx.doi.org/10.1007/978-981-10-5281-1_36.
- Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *9th international conference on learning representations, ICLR 2021, Virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: <https://openreview.net/forum?id=nzplWnVAYah>.

- Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. S. (2013). What yelp fake review filter might be doing? In E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, & I. Soboroff (Eds.), *Proceedings of the seventh international conference on weblogs and social media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. The AAAI Press, URL: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006>.
- Mulder, W. D., Bethard, S., & Moens, M. (2015). A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language*, 30(1), 61–98. <http://dx.doi.org/10.1016/j.csl.2014.09.005>.
- Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. In L. Vanderwende, H. D. I.I.I., & K. Kirchhoff (Eds.), *Human language technologies: Conference of the North American chapter of the association of computational linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA* (pp. 497–501). The Association for Computational Linguistics, URL: <https://www.aclweb.org/anthology/N13-1053/>.
- Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *The 49th annual meeting of the association for computational linguistics: Human language technologies, proceedings of the conference, 19-24 June, 2011, Portland, Oregon, USA* (pp. 309–319). The Association for Computer Linguistics, URL: <https://www.aclweb.org/anthology/P11-1032/>.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015: Technical Report*.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. A. Walker, H. Ji, & A. Stent (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)* (pp. 2227–2237). Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/n18-1202>.
- Rayana, S., & Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, & G. Williams (Eds.), *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, NSW, Australia, August 10-13, 2015* (pp. 985–994). ACM, <http://dx.doi.org/10.1145/2783258.2783370>.
- Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385, 213–224. <http://dx.doi.org/10.1016/j.ins.2017.01.015>.
- Santosh, K. C., & Mukherjee, A. (2016). On the temporal dynamics of opinion spamming: Case studies on yelp. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, & B. Y. Zhao (Eds.), *Proceedings of the 25th international conference on world wide web, WWW 2016, Montreal, Canada, April 11 - 15, 2016* (pp. 369–379). ACM, <http://dx.doi.org/10.1145/2872427.2883087>.
- Sechidis, K., Tsoumakas, G., & Vlahavas, I. P. (2011). On the stratification of multi-label data. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Lecture notes in computer science: vol. 6913, Machine learning and knowledge discovery in databases - European conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III* (pp. 145–158). Springer, http://dx.doi.org/10.1007/978-3-642-23808-6_10.
- Shang, Y., Liu, M., Zhao, T., & Zhou, J. (2021). T-bert: A spam review detection model combining group intelligence and personalized sentiment information. In I. Farkas, P. Masulli, S. Otte, & S. Wermter (Eds.), *Lecture notes in computer science: vol. 12895, Artificial neural networks and machine learning - ICANN 2021 - 30th international conference on artificial neural networks, Bratislava, Slovakia, September 14-17, 2021, Proceedings, Part V* (pp. 409–421). Springer, http://dx.doi.org/10.1007/978-3-030-86383-8_33.
- Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1), <http://dx.doi.org/10.1038/s41467-018-06930-7>.
- Shehnepoor, S., Salehi, M., Farahbakhsh, R., & Crespi, N. (2017). NetSpam: A network-based spam detection framework for reviews in online social media. *IEEE Transactions on Information Forensics and Security*, 12(7), 1585–1595. <http://dx.doi.org/10.1109/TIFS.2017.2675361>.
- Shehnepoor, S., Togneri, R., Liu, W., & Bennamoun, M. (2020). GANGster: A fraud review detector based on regulated GAN with data augmentation. CoRR abs/2006.06561. URL: <https://arxiv.org/abs/2006.06561>, arXiv:2006.06561.
- Stanton, G., & Irissappane, A. A. (2019). GANs for semi-supervised opinion spam detection. In S. Kraus (Ed.), *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, Macao, China, August 10-16, 2019* (pp. 5204–5210). ijcai.org, <http://dx.doi.org/10.24963/ijcai.2019/723>.
- Tang, T., Tang, X., & Yuan, T. (2020). Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8, 193248–193256. <http://dx.doi.org/10.1109/ACCESS.2020.3030468>.
- Tao, J., & Fang, X. (2020). Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1), 1. <http://dx.doi.org/10.1186/s40537-019-0278-0>.
- Tsoumakas, G., Katakis, I., & Vlahavas, I. P. (2010). Mining multi-label data. In O. Maimon, & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (2nd ed.). (pp. 667–685). Springer, http://dx.doi.org/10.1007/978-0-387-09823-4_34.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008). URL: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- Wu, G., Greene, D., Smyth, B., & Cunningham, P. (2010). Distortion as a validation criterion in the identification of suspicious reviews. In C. L. Giles, P. Mitra, I. Perisic, J. Yen, & H. Zhang (Eds.), *Proceedings of the 3rd workshop on social network mining and analysis, SNAKDD 2009, Paris, France, June 28, 2009* (pp. 10–13). ACM, <http://dx.doi.org/10.1145/1964858.1964860>.
- 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., ... 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.
- You, Z., Qian, T., & Liu, B. (2018). An attribute enhanced domain adaptive model for cold-start spam review detection. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018* (pp. 1884–1895). Association for Computational Linguistics, URL: <https://aclanthology.org/C18-1160/>.
- Zaeem, R. N., Li, C., & Barber, K. S. (2020). On sentiment of online fake news. In M. Atzmueller, M. Coscia, & R. Missaoui (Eds.), *IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM 2020, the Hague, Netherlands, December 7-10, 2020* (pp. 760–767). IEEE, <http://dx.doi.org/10.1109/ASONAM49781.2020.9381323>.
- Zhang, W., Du, Y., Yoshida, T., & Wang, Q. (2018). DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing and Management*, 54(4), 576–592. <http://dx.doi.org/10.1016/j.ipm.2018.03.007>.
- Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., & Cui, L. (2020). BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. In *2020 international joint conference on neural networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020* (pp. 1–8). IEEE, <http://dx.doi.org/10.1109/IJCNN48605.2020.9206973>.
- Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2021). Revisiting few-sample BERT fine-tuning. In *9th international conference on learning representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, URL: <https://openreview.net/forum?id=c0IH43yUF>.
- Zhang, M., & Zhou, Z. (2006). Multi-label neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18(10), 1338–1351. <http://dx.doi.org/10.1109/TKDE.2006.162>.
- Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), 109:1–109:40. <http://dx.doi.org/10.1145/3395046>.