IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 April 2023, accepted 8 May 2023, date of publication 18 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3277490

RESEARCH ARTICLE

A New Italian Cultural Heritage Data Set: Detecting Fake Reviews With BERT and ELECTRA Leveraging the Sentiment

ROSARIO CATELLI[1], LUCA BEVILACQUA[2], NICOLA MARINIELLO[2], VLADIMIRO SCOTTO DI CARLO[2], MASSIMO MAGALDI[2], HAMIDO FUJITA[3,4,5], (Life Senior Member, IEEE), GIUSEPPE DE PIETRO[1], AND MASSIMO ESPOSITO[1]

[1]Institute for High Performance Computing and Networking (ICAR), National Research Council, 80131 Naples, Italy
[2]Engineering Ingegneria Informatica S.p.A., 80142 Naples, Italy
[3]Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City 7000, Vietnam
[4]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18012 Granada, Spain
[5]Faculty of Software and Information Science, Iwate Prefectural University, Takizawa, Iwate 020-0611, Japan

Corresponding author: Rosario Catelli (rosario.catelli@icar.cnr.it)

ABSTRACT The growth of the online review phenomenon, which has expanded from specialised trade magazines to end users via online platforms, has also increasingly involved the cultural heritage of countries, a source of tourism and growth driver of local economies. Unfortunately, this has been paralleled by the emergence and spread of the phenomenon of fake reviews, against which the scientific world has developed language models capable of distinguishing them from the truthful. The application of such models, often based on deep neural networks with transformer-type architectures, is however limited by the availability of local language data sets for specific domains, useful for both training and verification. The purpose of this article is twofold. Firstly, a new data set was created in the Italian language, generally considered low-resource, relating to the domain of cultural heritage in Italy, by collecting reviews available online, reorganising them in the form of a data set usable by the language models. Secondly, a baseline of results for the detection of misleading reviews was constructed by exploiting two widely used language models, namely BERT and ELECTRA. The performance achieved is interesting, around 95% accuracy and F1 score, using data set splits between training and testing of 80/20 and 90/10. In addition, SHAP was used as a tool to support the explicability of AI models: in this way, it was possible to show the usefulness of sentiment analysis as a support for the recognition of deceptiveness.

INDEX TERMS Italian cultural heritage, data set, fake reviews, sentiment analysis, deceptive.

I. INTRODUCTION

The artistic and cultural heritage of every country needs substantial resources to be safeguarded from neglect. In order to find the economic resources necessary for this purpose, guided tours and exhibitions are often organized to attract tourists willing to pay both for the services offered and to enjoy the beauty and history offered by the places visited. However, the mere staging of such events is not enough to attract a demanding and not occasional public: the growing presence of websites and mobile applications that allow people to leave an opinion about the services offered and the state of the places, makes it increasingly necessary to prepare a targeted offer ready to improve, also to face the rise of competition due to ranking platforms and their algorithms [1].

In this sense, the exploitation of artificial intelligence techniques leaves ample room for improvement [2], [3]. In particular, the field of Natural Language Processing (NLP) can

The associate editor coordinating the review of this manuscript and approving it for publication was Sergio Consoli.

help to distinguish genuine reviews from fake ones by correctly directing efforts to modernize and improve the services offered for cultural heritage. From a technical point of view there is a wide availability of useful tools that take advantage of machine learning techniques, on the other hand there is a strong limitation due to the language of the analysis or the available data sets: most of the data sets are in English and it often becomes difficult to obtain satisfactory performance for different languages, often indicated as low-resource languages, such as Italian. For example, the literature provides several review data sets in English language, which have as object restaurants, hotels or doctors [4], [5], [6], or even products purchased online [7] or mobile applications [8].

In this paper a new data set in Italian language was proposed, as far as it is known unique in its kind, which has as object of the reviews several cultural places of interest and for which there is also an indication related to the polarity of the sentiment, that is positive or negative: its content is built taking reviews from several web pages. In order to build a valid baseline that could be useful for future comparison, this data set was tested with two of the latest and most popular classification systems already employed in other domains to distinguish deceptive and genuine reviews, namely BERT and ELECTRA. In addition, sentiment information was exploited, trying to understand to what extent it can help in distinguishing true and fake reviews.

The rest of this article is structured as follows: section II provides the information needed to understand the state of the art in the literature related to what is proposed, while section III gives information related to both the released data set and the language model used to provide a baseline of performance in the cultural heritage domain; instead, sections IV and V analyzes and discusses the results obtained, then section VI reaches conclusions and suggests possible directions for future work.

II. BACKGROUND AND RELATED WORKS

In this section, an overall outlook is provided with regard to the state of the scientific literature underlying the proposed article: specifically, section II-A provides a brief view with regard to the application of the latest computing techniques to the field of cultural heritage, while section II-B supplies information related to the specific task focused on for this paper, thus the NLP techniques applied. In addition, section II-C provides details on problem transformation methods for dealing with multi-label problems.

A. CULTURAL HERITAGE DOMAIN: CURRENT APPROACHES

In the field of cultural heritage, the use of Machine Learning (ML) and Deep Learning (DL) techniques is still quite limited today: as a rule, often also due to the non-computer science background of the personnel employed in this field, the use of ready-to-use toolboxes based on statistical methods is preferred, not providing any feedback to the scientific world on the ML and DL sides. In fact, the thrust coming from the cultural heritage sector related to ML and DL is predominantly on computer vision, than on other areas such as the one covered in this paper, i.e. NLP, that is still limited [9], [10]. In particular, the lack of adequate training datasets further invalidates the possibilities of using the latest techniques, whether supervised, semi-supervised or unsupervised. In fact, it is easier to find application of techniques such as linear and logistic regression, support vector machines and so on [11], although the progressive digitization of historical books and texts will bring about a necessary progression toward language models based on deep neural networks, finally having larger text corpora available to proceed with training. In this regard, at present, increasing the amount of data sets available especially in low-resource languages is a necessary step: although the available pre-trained language models are not created for the purpose of handling information related to the cultural heritage domain, the availability of labeled data sets albeit of modest size finally allows the application of techniques to fine-tune such models to the specific task.

For these reasons, this paper proposes both a new Italian-language dataset dedicated to the cultural heritage sector and a baseline of the performance achievable by employing state-of-the-art language models.

B. FAKE REVIEW CLASSIFICATION SYSTEMS

The recent proliferation of the review system, as a viaticum for user evaluation in the quality of a service, has led to the rapid rise of fraudulent mechanisms aimed at promoting or denigrating the services and/or product offered, through bogus reviews designed to alter perceived quality. From a scientific point of view, the detection of reviews created ad hoc for such purposes is a major technological challenge: even the development of automated systems is held back by the difficulty in finding data sets for training, caused by the economic motivations of platforms operating in the field. Therefore, the lack of reliable data leaves ample room for maneuver for spammers who remain difficult to detect.

Fortunately, the release of large data sets related to platforms such as Amazon or Yelp has nonetheless allowed experimentation with a variety of techniques, albeit with a focus on the online sales sector. Reference [12] have provided an in-depth look at this issue, considering both traditional machine learning approaches based on statistical techniques (e.g., Ensemble systems, Support Vector Machines, Random Forest or Naive Bayes) and deep learning approaches based on deep neural networks (convolutional, recurrent or generative). An important distinction made in the literature separates systems into two macrocategories based on whether or not they use additional information about reviewers [13], such as their social profiles: such content is often maintained by companies unwilling to release it freely, partly for privacy reasons. Therefore, the focus of the proposed approach will hinge on being able to take full advantage of reviews alone as the primary information i.e., by relying on linguistic features,

employing language models based on deep neural networks such as BERT and ELECTRA.

The creation of a data set for fine-tuning supervised machine learning systems has always been a cause for discussion in the literature: duplicating or copying and modifying the collected reviews to create the bogus versions [14] has been a procedure criticized by many because of the lack of reliability [5], [15], often replaced by the practice of paying anonymous workers to obtain bogus pseudo-reviews [5], [6] but also criticized for the lack of veracity of the reviews thus obtained, both because of the lack of knowledge of the domain and the different psychological state and experience of the workers hired for the purpose [4]. Basically, there is to date no one procedure considered superior to the others in terms of the quality of the data set obtained.

Although some attempts have been made with semi-supervised [16], [17] and unsupervised [4], [18], [19] to overcome the problem of labeling data sets, more extensive results in the literature have employed supervised methods such as Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Multilayer Perceptron (MLP) or hybrid [15], [20], [21], [22], or even based on K-Nearest Neighbors (k-NN) [23].

The proposed approach aims to provide a baseline with the latest language model-based methods, even for the Italian-language cultural heritage domain, exploiting the integration of information content due to sentiment in the fake review recognition process.

1) SENTIMENT-AWARE SYSTEMS

Experimentation carried out over the past decade in the research community made it clear how sentiment affected the characteristics of writing and, consequently, of reviews. In particular, the predisposition to deception on the part of writers showed a number of specific patterns: already [24] had shown a greater bias toward extreme sentiment in deceptive reviews than in truthful ones, and this was later confirmed by [6] and [25], which suggested how a shortage or abundance of first-person singular pronouns respectively represented a way to distance oneself from the fake or make physical presence more credible in, for example, a place.

Earlier attempts to integrate sentiment information within the systems for detecting deceptive reviews [15], [26] had already shown improvements by exploiting different features of texts, from lexical patterns to syntactic stylometry. References [21] and [27] tried to formalize the contemporary results by summarizing four key characteristics of texts, namely comprehensibility, informativeness, writing style, and cognitive indicators, as the key points to focus on in order to distinguish authentic and manipulated reviews according to the positive or negative sentiment expressed. Instead, [28] have tried to exploit emotions as a discriminator between true and misleading opinions without obtaining significant results. In addition, [29] have tried to construct two independent components, one dedicated to misleadingness and one to sentiment, for detecting bogus negative sentiment reviews. Reference [30] additionally showed how the use of sentiment could help to obtain classifiers with lower bias, while [31] showed the presence of correlation of positive sentiment and truthfulness on the one hand and negative sentiment and deceptiveness on the other. More recently, [32] employed hybrid machine learning and deep learning models associated with external vocabularies to provide numerical sentiment evaluation of reviews, while [33] combined sentiment information from an additional external labeling process into BERT embeddings.

What emerges from a review of the recent literature is how the more recent linguistic models, such as BERT and ELECTRA, have been little used in the context of integrating the information content from sentiment analysis with the deceptiveness characteristics of the text, even more so when dropped into the context of cultural heritage, although its potential in the use of sentiment has been glimpsed [34], [35], [36], [37].

C. MULTIPLE LABELS: PROBLEM TRANSFORMATION METHODS

As mentioned earlier, the growth of the rating system through reviews in recent years has been market-driven: when thinking about streaming services, it is evident how the need to provide suggestions that are increasingly in line with user needs has required a refinement of multi-label rating systems to identify the different genres in which to frame the products offered, such as in music [38], [39] or movie [40], [41] streaming services.

Said $l \in L$ ($|L| > 1$) to be the single label associated with the example to be classified, if the set of disjoint labels L is equal to or greater than 2 a binary or multi-class classification problem occurs, respectively. In the multi-label case each example is no longer associated with l but with the set $Y \subseteq L$, with L no longer a disjoint set. Already with $|L| = 2$, four combinations are possible therefore four possible classifications of the example: the concepts of binary and multi-class classification are lost because the classes are no longer mutually exclusive [42].

Therefore, multi-label problems have been addressed through problem transformation methods: in particular, the most basic, called Binary Relevance (BR) methods, have consisted of transforming to multiple binary problems with as many binary classifiers for each label or to multiple multi-class problems but in each case showing limitations due to the inability to model dependencies between labels. The consequent evolution due to the need to overcome the limitation consisted of the Chain Classifier method, in which single-label classifiers are cascaded in such a way as to have view of both the example to be classified and the result provided for a different label by the previous binary classifier. Finally, a further approach, to which the work proposed in this paper draws on, is the Label Powerset (LP) method, which transforms the multi-label problem into a single-class,

single-label problem by defining the powerset as a set of 2^L possible combinations: this takes into account the dependencies between labels and allows for better performance than the BR and Chain Classifier methods, but it runs into computational complexity that grows exponentially with L. In particular, assuming $[n, m]$ the pair of labels related to deceptiveness and sentiment polarity and thus being $L = 2$ then $|P| = 4$ occurs.

Other more complex methods, called algorithm adaptation such as Multi-Label k-NN or Instance-Based Learning based on LR or Back-Propagation for Multi-Label Learning (BP-MLL), have tried to solve such problems in their own space without simplifying them: by employing modern deep learning-based language models such as BERT and ELECTRA, it is possible to automatically manage the feature space in the multi-label domain and, in the case of the proposed study, to understand the extent to which sentiment information can help improve performance in recognizing misleading reviews.

III. MATERIAL AND METHODS

This section introduces the proposed Italian Cultural Heritage data set in section III-A, while section III-B introduces the architectures of the language models used to build the provided baseline.

A. ITALIAN CULTURAL HERITAGE (ICH) DATA SET

The creation of the Italian Cultural Heritage (ICH) data set draws inspiration from the data set *Deceptive Opinion Spam Corpus v1.4* [1] (DOSC) [5], [6], from which the scope of the reviews and the size change: in the latter case, reviews were collected for 20 hotels in Chicago, for a total of 1600 reviews perfectly distributed between true, fake, with both positive and negative sentiment; in the case of the ICH data set it refers to the scope of Italian cultural heritage, particularly those of the city of Naples, and there are 800 reviews divided as shown in Table 1. Specifically, 10 positive and 10 negative reviews were collected for each of the 20 chosen places in the city of Naples (in Italian *Napoli*), namely:

1) Napoli Sotterranea
2) Cappella Museo Sansevero
3) Galleria Borbonica
4) Catacombe di San Gennaro
5) Museo Archeologico Nazionale di Napoli
6) Teatro di San Carlo
7) Castel Sant'Elmo
8) Palazzo Reale
9) Cimitero delle Fontanelle
10) Reggia di Carditello - Real Sito di Carditello
11) Via San Gregorio Armeno
12) Lungomare Caracciolo
13) Spaccanapoli
14) Castel dell'Ovo
15) Complesso Monumentale di Santa Chiara
16) Duomo di Napoli
17) Piazza del Plebiscito
18) Chiesa del Gesù Nuovo
19) Toledo (Art Metro Station)
20) Porta Nolana

TABLE 1. Italian Cultural Heritage dataset composition.

# Reviews	Negative	Positive
Deceptive	200	200
Truthful	200	200

In addition, the average length of the collected reviews is about 61 words per review.

In the case of the DOSC data set, truthful reviews were collected from different platforms such as TripAdvisor, Expedia, Hotels.com, Orbitz, Priceline, and Yelp, considering positive those with 5 stars and negative those with 1 or 2 stars. In contrast, in the case of the ICH data set, the reviews considered true were collected exclusively from the TripAdvisor platform, considering positive those with a 4 or 5 star rating while negative those with 1 or 2 stars. Regarding the fake reviews, those in the DOSC data set were collected through Amazon Mechanical Turk[2] (AMT) crowd-sourcing service, while in the case of the ICH data set, a different approach was preferred, constructing the fake reviews from the true ones that were appropriately modified.

Creating a data set of fake reviews is generally a complex task and exposes itself to several criticisms however one proceeds, so it needs clarification. This is because, generally, the peculiar case of data sets of fake reviews suffers from the problem of collecting and distinguishing into true and false the basic constituent element of the data set itself, namely the review.

The very fact of considering the review acquired by a site such as TripAdvisor as true may not be properly correct and vitiated by the presence of a filtering mechanism, proprietary to the platform. For obvious reasons of intellectual property protection, such filtering algorithms are not generally available to the scientific community for analysis but, even by virtue of the commercial reasons behind them, they can be considered bona fide since their goal is to protect both the good name of the platform itself and those who read by trusting it.

Similarly, constructing the misleading part of such a data set can be done in different ways, all of which can be criticized for different aspects. In the case of the DOSC data set, the authors created 1600 Human Intelligence Tasks in the AMT platform, offering a $1 reward for *Turks* residing in the US, providing two constraints, namely 30 minutes to complete the task and only one author per review (to avoid confusing classifiers with different writing styles): the *Turks* pretended to work for or visit the indicated hotel then wrote the review while the authors flunked reviews found to be too short or

[1] https://myleott.com/op-spam. html

[2] https://www.mturk.com/

copied. In contrast, as mentioned above, in the case of the ICH dataset it was decided to proceed differently: in particular, reviews that were considered to be true were given to 3 practitioners working in the cultural heritage sector and they were asked to work cooperatively on each review in order to construct a similar one that expressed through some variation (e.g. of vocabulary or by substituting original phrases with invented ones) a similar sentiment to the original one. For example:

> *La cooperativa onlus La Paranza ha dato vita a questo sogno: dare dignità al quartiere Sanità, portando tanti turisti grazie alle Catacombe. Ci sono riusciti! Anche solo per questi ragazzi e per il loro sogno vale la pena visitarle. Noi siamo stati fortunati perchè alla visita era associato un aperitivo a base di birra e taralli molto buoni! Le catacombe sono tenute molto bene e le spiegazioni sono state interessanti ed esaustive. Molto brava la guida!*
> English version: *The non-profit cooperative La Paranza gave life to this dream: to give dignity to the Sanità neighborhood, bringing many tourists thanks to the Catacombs. They have succeeded! If only for these guys and their dream, they are worth visiting. We were lucky because the visit was associated with an aperitif of beer and very good taralli! The catacombs are very well kept and the explanations were interesting and comprehensive. Very good the guide!*

This sample review is part of the set of true positives: in principle, for the reader, some detail regarding the information of the lived experience shines through. If, on the other hand, one reads the bogus positive version:

> *Le catacombe di San Gennaro sono di gran lunga uno dei migliori siti culturali che io e la mia famiglia abbiamo avuto il piacere di visitare durante il nostro soggiorno a Napoli. Il servizio di guida nel sito è il migliore che si possa desiderare, per non parlare di quello che c'è da vedere che è da morire. Questo è un ottimo posto per portare la famiglia a vivere un'avventura in un luogo misterioso e antico. Consiglierei questo luogo a chiunque. 5 stelle!*
> English version: *The Catacombs of San Gennaro are by far one of the best cultural sites my family and I have had the pleasure of visiting during our stay in Naples. The guide service at the site is the best you could ask for, not to mention what there is to see is to die for. This is a great place to take the family on an adventure to a mysterious and ancient place. I would recommend this place to anyone. 5 stars!*

What emerges is a general description, not too detailed, which is usually a harbinger of deceptiveness, as well as not exclusively employing the first person singular but trying to change the subject, trying to *shift the blame for the false* at least from a psychological point of view, and thus distancing oneself from what one is writing. This situation, which has already been extensively analyzed in the literature [6], is one of the features that classification systems based on deep learning techniques, albeit in their complex dimensional space, can take into account in order to effectively distinguish true and fake reviews. Similarly with regard to sentiment, exaggeration is often a symptom of falsehood [24], [25], hence the possibility of exploiting sentiment analysis as an advantage.

TABLE 2. Hyper-parameters of $\text{BERT}_{\text{BASE}}$ based models.

Hyper-parameter	Value
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

B. ARCHITECTURES

In the following, a general description of the BERT architectures in Section III-B1 and ELECTRA in Section III-B2 is provided, while Section III-B3 provides some implementation details specific to this case study and common to the two architectures.

1) BERT

Although the field of cultural heritage is highly specialized, if not niche from the point of view of the NLP field, the development of the latest language models based on deep learning techniques has made it possible to easily adapt and employ such tools in the most diverse domains. *Bidirectional Encoder Representations from Transformers*, also known as BERT and proposed by [43], is an architecture based on transformer-type deep neural networks [44], of which the generally most widely used model is provided in a pre-trained form, that is, with a pre-established linguistic knowledge base thanks to training carried out on huge text corpora managed by processors with high computational capacity. Thanks to this, it is then possible to start from the pre-trained model and carry out what is called fine-tuning, that is, specializing the model on specific texts and tasks for its intended purpose, be it sentiment analysis, named entity recognition or any other NLP task. The advantage lies in the ability of the model to retain prior knowledge in the innermost neural network layers by generalizing them, so by replacing or varying the outermost layers it is possible to specialize it flexibly without losing pre-existing information. In particular, in order to carry out fine-tuning, it is necessary to employ a number of hyper parameters whose values directly influence the results that can be obtained: the main ones are given directly in Table 2.

The architecture of BERT also involves the use of two particular tokens, [CLS] and [SEP]. The first is a vector of size H, i.e., the size of the hidden layers, which is obtained at the output and has the task of representing the entire sequence to be provided as input to a subsequent arbitrary classifier. The second one performs the task of a separator between sentences, with use varying according to specific tasks. In addition, the use of the WordPieceModel-based tokenizer [45], which divides words into common subwords, allows better handling of the out-of-vocabulary (OOV) word problem, as well as also reducing the size of the vocabulary itself.

The input of the final classification layer is provided by the last hidden layer and denoted as a vector $C \in R^H$. Provided instead the parameter matrix of the classification layer $W \in R^{KxH}$ with K number of categories, then the probability P for each category is given by:

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

a: TRANSFORMER

[44] introduced the transformer,the most important component of the BERT architecture. As introduced above, 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:

$$\hat{h}_i^0 = E(x_i) + E(i) + E(1_{\mathbf{x}}) \tag{2}$$

$$\hat{h}^0_{\mathrm{j+|x|}} = E(y_j) + E(j + |x|) + E(1_{\mathbf{y}}) \tag{3}$$

$$\hat{h}^0_{\cdot} = Dropout(LN(\hat{h}^0_{\cdot})) \tag{4}$$

Therefore the embedding encountersat first M transformer blocks where, considered FF the Feed Forward layer and $GELU$ the element-wise Gaussian Error Linear Units activation function [46] while MHSA is the Multi-Heads Self-Attention function respectively, it results:

$$\hat{h}.^{\mathrm{i+1}} = Skip(FF, Skip(MHSA, h.^i)) \tag{5}$$

$$Skip(f, h) = LN(h + Dropout(f(h))) \tag{6}$$

$$FF(h) = GELU(h\mathbf{W}_1{}^\top + \mathbf{b}_1)\mathbf{W}_2{}^\top + \mathbf{b}_2 \tag{7}$$

where $h^i \in \mathbb{R}^{(|\mathbf{x}|+|\mathbf{y}|)\times \mathrm{d_h}}$, $\mathbf{W}_1 \in \mathbb{R}^{4\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_1 \in \mathbb{R}^{4\mathrm{d_h}}$, $\mathbf{W}_2 \in \mathbb{R}^{4\mathrm{d_h}\times \mathrm{d_h}}$, $\mathbf{b}_2 \in \mathbb{R}^{4\mathrm{d_h}}$ and the new $\hat{h}_i$ position is:

$$\begin{aligned}[\ldots, \hat{h}_i, \ldots] &= MHSA([h_1, \ldots, h_{|\mathrm{x}|+|\mathrm{y}|}]) \\ &= \mathbf{W}_o Concat(h_i^1, \ldots, h_i^N) + \mathbf{b}_o\end{aligned} \tag{8}$$

In each attention head results:

$$h_i^j = \sum_{k=1}^{|\mathbf{x}|+|\mathbf{y}|} Dropout(\alpha_k^{(i,j)})\mathbf{W}_V^j h_k \tag{9}$$

$$a_k^{(i,j)} = \frac{exp\frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_k}{\sqrt{d_h/N}}}{\sum_{k'=1}^{|\mathbf{x}|+|\mathbf{y}|} exp\frac{(\mathbf{W}_Q^j h_i)^\top \mathbf{W}_K^j h_{k'}}{\sqrt{d_h/N}}} \tag{10}$$

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

2) ELECTRA

ELECTRA, that stands for *Efficiently Learning an Encoder that Classifies Token Replacements Accurately*, is an innovative method for learning language representation in a self-supervised and less computationally intensive way and was proposed by [47]. Based on transformer networks, its pre-training involves two transformer models called *Generator G* and *Discriminator D*: the former predicts [MASK] tokens and replaces them with fakes in the input sequence, the latter is trained to detect them. Specifically, G is trained to predict the original tokens that have been randomly replaced by a [MASK] token within the input sequences; after which G replaces the [MASK] tokens with fakes. Finally, D tries to predict whether the tokens are original or fake.

Given an input sequence s of tokens $s = w_1, w_2, \ldots, w_n$, where w_t $(1 \leq t \leq n)$ represents the generic token (e.g. character, sub-word or word), s is encoded into a sequence of contextualized vector representations $h(s) = h_1, h_2, \ldots, h_n$ by both G and D.

For each position t for which $w_t = [MASK]$, G outputs, through a softmax layer, the probability p to generate a specific token w_t:

$$p_G(w_t|s) = \frac{e(w_t)^T h_G(s)_t}{\sum_{w'} exp(e(w')^T h_G(s)_t)} \tag{11}$$

where $e(\cdot)$ is the embedding function $e(\cdot) : w_t \in s \longrightarrow \mathbb{R}^{dim}$ where dim is the chosen embedding size.

While D predicts whether w_t is original or not through a sigmoid layer:

$$D(s, t) = sigmoid(e(w_t)^T h_D(s)_t) \tag{12}$$

where $sigmoid(x) : x \in \mathbb{R}^N \longrightarrow [0, 1]$.

G leverages this loss function for the pre-training phase:

$$\mathcal{L}_{Gen} = \mathcal{L}_{MLM} = \mathbb{E}(\sum_{i \in m} -\log p_G(w_i|s^{masked})) \tag{13}$$

where s^{masked} is the sentence with the masked words and $m = m_1, m_2, \ldots, m_k$ are k random selected words.

Instead D uses this other loss function:

$$\begin{aligned}\mathcal{L}_{Dis} = \mathbb{E}(\sum_{t=1}^{n} &-\mathbb{I}(w_t^{corrupt} = x_t)\log D(s^{corrupt}, t) + \\ &- \mathbb{I}(w_t^{corrupt} \neq x_t)\log D(s^{corrupt}, t))\end{aligned} \tag{14}$$

where $s^{corrupt}$ is the altered sentence, while $w_t^{corrupt}$ is the altered word.

Finally, this combined loss is minimized:

$$\min_{\theta_G, \theta_D} \sum_{s \in \chi} \mathcal{L}_{Gen}(s, \theta_G) + \lambda \mathcal{L}_{Dis}(s, \theta_D) \tag{15}$$

In this way, predictions are computed for each token and the discriminator loss can be computed on all input tokens. This is the main reason for the higher efficiency of ELECTRA

compared to masked language models such as BERT, where the model loss is calculated only on the masked tokens.

G is exclusively used for the pre-training phase, while D is effectively used for fine-tuning on the specific task. As seen in [48], a dense layer which uses a softmax activation function is added to D, hence a cross-entropy loss function is used:

$$\mathcal{L} = \sum_{i=1}^{n} -\log\left(\frac{e^{s_{i,l_i}}}{\sum_{j=1}^{k} e^{s_{i,c_j}}}\right) \tag{16}$$

where: n is the total number of labeled tokens, l_i is the label of the i^{th} token, k is the number of categories for the different labels and c_j is any of these categories. Furthermore, s_{i,l_i} represents the score for the i^{th} token evaluated as belonging to its correct category l_i, while s_{i,c_j} represents the score for the same i^{th} token evaluated as belonging to the category c_j.

3) IMPLEMENTATION DETAILS

The pre-trained versions employed of BERT and ELECTRA in the Italian language are those provided by the MDZ Digital Library team of the Bavarian State Library through the Hugging Face framework.[3] Specifically, the provided versions come in a standard variant and an XXL variant (those employed), so the training corpora vary. In detail, while the former relies on text collected through a Wikipedia dump and various collections from OPUS[4] with 13 GB of final size and about 2 billion tokens included, the latter case adds text from the OSCAR corpus,[5] reaching 81 GB and 13 billion tokens included.

Such models, built on $\text{BERT}_{\text{BASE}}$, involve by default the use of a loss function of type *Categorical Cross Entropy*, which, however, is suitable for multiclass classification. In order to be able to combine, through the Label Powerset problem transformation method, the labels related to deceptiveness and sentiment polarity in order to exploit the advantages given by relative feature extraction, a loss function of type *Binary Cross Entropy* (BCE[6]) was employed. To improve numerical stability, use was made of a variant of it with Logits (BCEwL[7]), which achieves this result through combination with a sigmoid and the use of the LogSumExp function (LSE[8]). Given N the batch size, the BCEwL for single-label classification as employed in the proposed architectures can be described as:

$$\begin{aligned} l(x, y) &= L = \{l_1, \ldots, l_N\}^T, \\ l_n &= -w_n[y_n \cdot log\sigma(x_n) + (1 - y_n) \cdot log(1 - \sigma(x_n))] \end{aligned} \tag{17}$$

A simple flowchart (Figure 1) of the process has been included for completeness.

[3]https://huggingface.co/dbmdz/
[4]http://opus.nlpl.eu/
[5]https://traces1.inria.fr/oscar/
[6]https://pytorch.org/docs/stable/generated/torch.nn.BCE.html
[7]https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html
[8]https://en.wikipedia.org/wiki/LogSumExp

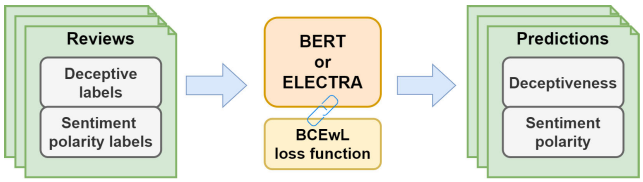


FIGURE 1. A simple flowchart of the process.

TABLE 3. BERT and ELECTRA results.

		Training 80% - Test 20%		Training 90% - Test 10%	
		BERT	ELECTRA	BERT	ELECTRA
Polarity	F_1	94.99	94.73	95.03	95.71
	A	95.05	94.87	95.10	95.80
Deceptiveness	F_1	97.70	97.27	97.31	96.63
	A	97.67	97.22	97.30	96.65
Both	F_{1M}	96.60	96.44	96.98	96.73
	A_M	93.32	92.92	94.05	93.45
	F_{1P}	95.20	95.57	96.47	96.10
	A_P	95.32	95.65	96.50	96.10
	F_{1D}	97.93	97.58	97.48	97.35
	A_D	97.90	97.57	97.45	97.30

IV. RESULTS

This section analyzes the results obtained from a quantitative point of view, referring in particular to the accuracy and F_1 score values obtained depending on the experimental setup.

Within Table 3 are the results obtained from the experiments performed. The results reported are the result of an experimental session in which each test case was repeated 5 times for each of the 5 different seeds used to shuffle the original dataset, for a total of 25 experiments per test case: those reported are the values obtained from the arithmetic mean of the results. This mode was repeated for a training set to test set ratio of both 90 to 10 and 80 to 20. The metrics considered for evaluating the results are accuracy, which provides a score on the basis of exact predictions relative to the total, and the F_1 score (harmonic mean of accuracy and recall), which takes into account not only exact predictions but also their equal distribution among the different classes. In addition, the experiments were conducted taking into account either deceptiveness alone, polarity alone or both labels combined as illustrated in section II-C and implemented using the BCEwL function as shown in section III-B3. For clarity, the result reported with subscript M constitutes the value given to the joint prediction of both labels, while the results reported with subscript P or D are the values of the specific predictions of polarity and deceptiveness, respectively, extracted from the experimental session of joint prediction of both labels which is indicated by the subscript M.

Looking at the results, there is a first general point to be made: both architectures benefit from the simultaneous use of the two labels. In fact, in both cases, there is a performance improvement in moving from the classical single-label prediction mode to the joint prediction mode using the proposed

methodology, and this supports the goodness of the method described in the article. In particular, BERT benefits from improvements ranging in the range of 0.15% to 0.27%, except for polarity in the 90/10 case where the improvement is much more pronounced and is around 1.42%. The ELECTRA case history, on the other hand, shows a fluctuating trend, with higher average improvements contained between 0.30% and 0.39% or between 0.65% and 0.84%.

A further look allows us to verify that in both training set to test set ratio scenarios, the BERT architecture succeeds in achieving better results than ELECTRA when referring to prediction on deceptiveness, while the scenario is more complex in relation to polarity. In fact, in a single prediction scenario, BERT is better when training set to test set ratio is 80/20 and ELECTRA when training set to test set ratio is 90/10. While moving to the joint prediction scenario basically a reversal of results occurs, whereby ELECTRA outperforms BERT when training set to test set ratio is 80/20 and is outperformed by BERT when training set to test set ratio is 90/10: it is likely that the different architectural structure affects BERT and ELECTRA to different degrees depending both on the proportion of data between training and test set and on the number of labels jointly predicted, and this could be the subject of further future analysis following further extension of the data set.

V. DISCUSSION

This section analyzes the results obtained from a qualitative point of view, describing some peculiar case histories that provide insights into the plausible motivations underlying the results obtained from the tested language models.

Qualitative analysis was carried out using the AI explainability tool called SHAP [49]: the approach of this tool is to perturb the inputs of a model and observe the changes in the output, so as to better understand the contributions made by a specific feature, and is essentially based on the calculation of Shapley values from coalitional game theory. Specifically, this technique assumes that (1) the payoff is represented by the prediction, (2) the players in each team are the feature values of a data instance, and (3) the Shapley values are those parameters that allow the payoff to be redistributed equally among the features. Applying this methodology to text, it evaluates the impact of the component text fragments of an input sentence, i.e. the features, on the prediction obtained from the language models also based on Transformers architecture [50].

Therefore, employing SHAP provides valuable support for the purposes of interpreting the possible reasons why certain reviews within the validation dataset are more or less easily recognized by BERT or ELECTRA. In addition, for the purpose of making the analysis more straightforward, it was preferred to look at models trained on 90 percent of the dataset, so as to have a greater focus on the most challenging texts to recognize. The results provided by this tool, in the form of a bar graph, attempt to attribute to the various constituent tokens of the sentence (as identified by the specific tokenizer of the language model) a defined score so as to try to specify their weight in the overall score given to the sentence, i.e., the ''probability'' that it is truthful or deceptive, keeping in mind that in this case the positivity of the score indicates the deceptiveness of the sentence, while the negativity of the score indicates truthfulness.

Furthermore, it is important to remember some useful aspects of performing such an analysis, as 5 different runs were performed for each of the 5 seeds used. First, it becomes necessary to select a unique seed so as to fix the split of dataset to be analyzed. Secondly, the results of all five runs performed are reported by averaging in order to have an unambiguous value related to the recognition ability for the specific phrase: it is clear that in this case, having performed 5 runs, the possible percentages are 20, 40, 60, 80 or 100%. Finally, the variability inevitably associated with the choice of the model then employed in SHAP: while it is true that the model chosen is the best obtained from that experimental session, it is equally impossible to predict whether the model selected is stronger or weaker in recognizing some phrases rather than others, as the best model chosen is downstream of the seed variation.

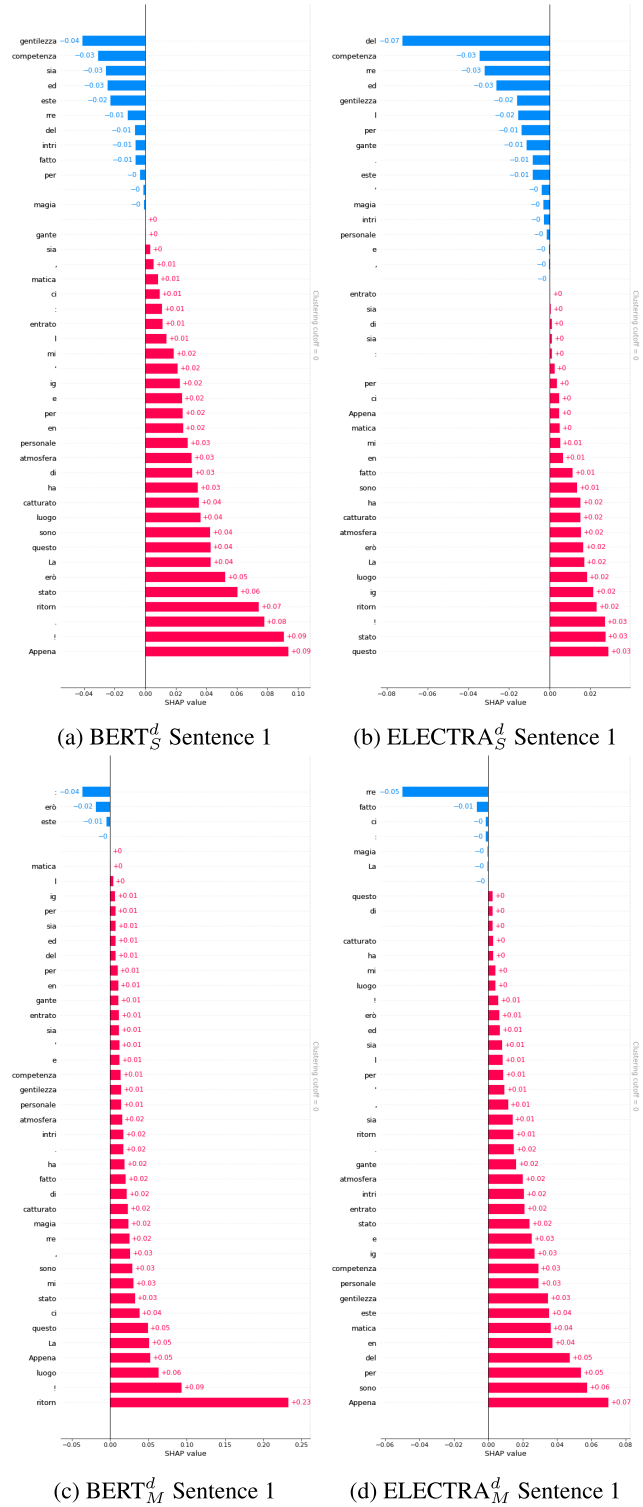
For example, the following Sentences 1 and 2 are taken as example and the related Figures 2 and 3 show the results provided by SHAP, employing BERT and ELECTRA in singlelabel recognition versions on deceptiveness and multilabel with focus on deceptiveness, denoted by superscript d and subscripts S and M respectively.

> *Appena entrato sono stato esterrefatto, sia per gentilezza e competenza del personale sia per l'atmosfera intrigante ed enigmatica. La magia di questo luogo mi ha catturato: ci ritornerò!*
> (English translation: *As soon as I entered I was stunned, both by the kindness and competence of the staff and by the intriguing and enigmatic atmosphere. The magic of this place captured me: I will return!*)

> *Dopo la recente passeggiata a spaccanapoli, posso sicuramente dire che non ci tornerò. Non offre nessuna esperienza in particolare e non c'è nulla da vedere, nemmeno negozi o botteghe. Vorrei dire a tutti che non conviene andarci perchè è tempo sprecato.*
> (English translation: *After the recent walk in spaccanapoli, I can definitely say that I will not go back there. It offers no experience in particular and there is nothing to see, not even stores or boutiques. I would like to tell everyone that it is not worth going there because it is wasted time.*)

The veracity of the proposed reviews is difficult to identify for all models (they should be true, instead they are identified as deceptive), however, two different behaviors can be distinguished.

In the case related to Sentence 1, the use of multilabel models tends to ensure that there is a further deviation toward

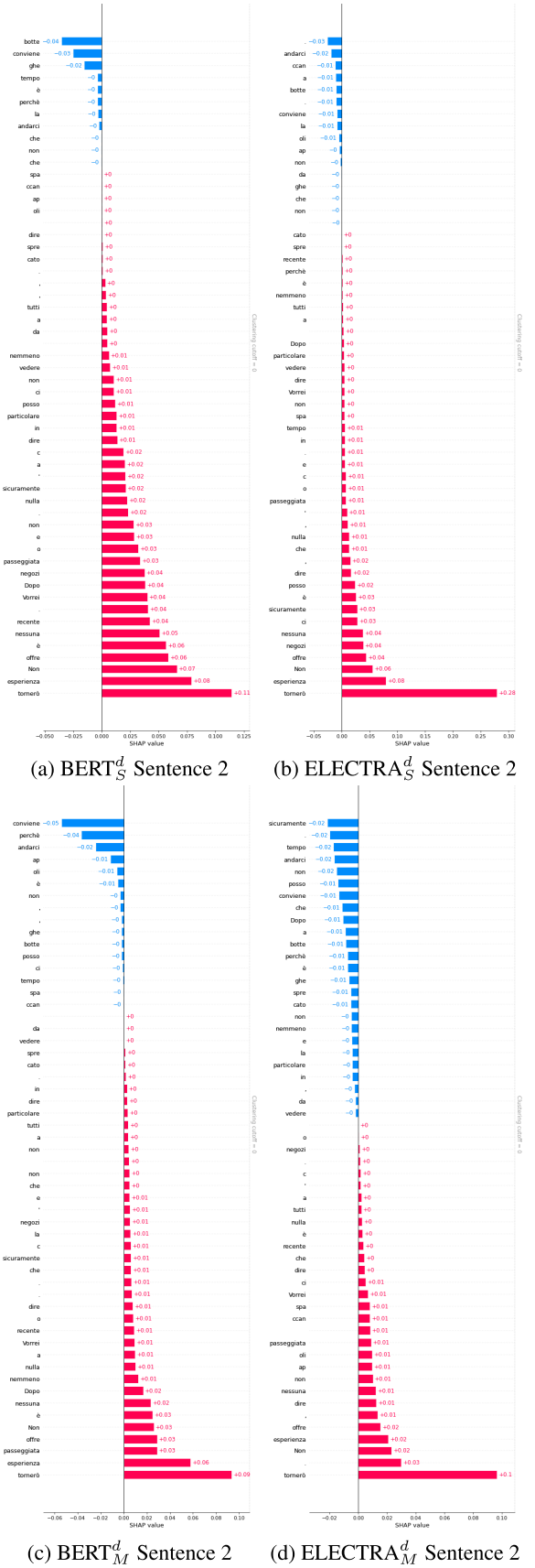


(a) BERT^d_S Sentence 1 (b) ELECTRA^d_S Sentence 1

(c) BERT^d_M Sentence 1 (d) ELECTRA^d_M Sentence 1

FIGURE 2. **Sentence 1.**

deceptiveness. In particular, it is interesting to note that in this case the focus is on two weight words, namely *competenza* (English translation: *competence*) and *gentilezza* (English translation: *kindness*), which move from scores of −0.03 and



(a) BERT^d_S Sentence 2 (b) ELECTRA^d_S Sentence 2

(c) BERT^d_M Sentence 2 (d) ELECTRA^d_M Sentence 2

FIGURE 3. **Sentence 2.**

−0.04 (BERT) or −0.03 and −0.02 (ELECTRA) to +0.01 (both in BERT) and +0.03 (both in ELECTRA), respectively. In addition, there is an overall redistribution of scores among the tokens, but this leads to a further expansion toward deceptiveness: in fact, truthfulness increases from −0.19 to −0.07 (BERT) and −0.24 to −0.06 (ELECTRA), i.e. it decreases, while deceptiveness increases from +0.94 to +1.02 (BERT) and +0.29 to +0.72 (ELECTRA), i.e. it increases.

Differently, in Sentence 2, the models veer toward truthfulness. Again, some more weighty words, such as the interrogative adverb *perchè* (English translation: *because*) and the modifier *sicuramente* (English translation: *definitely*) that can fulfill an adverb or interjection function for BERT and ELECTRA respectively, tend to change scores, going from 0 to −0.04 and +0.03 to −0.02. The redistribution of scoring that takes place among the tokens is different this time: both BERT and ELECTRA see decreases in deceptiveness scores, which fall from +0.95 and +0.80 to +0.47 and +0.31, and increases in truthfulness scores, which instead rise from −0.09 and −0.12 to −0.14 and −0.21 respectively.

This behavior highlights an important aspect: the use of sentiment as a supporting label in a mechanism such as the one proposed contributes to improving recognition ability by going to condition the outcome. In other words, the model chooses one class or the other with greater certainty.

Although the model behavior might appear to be independent of whether the result is correct or not, a closer analysis brings out a determining factor involving weight tokens. Such tokens drive the turn from a grammatical point of view; in particular, it is possible to infer a difference in the parts of speech involved in the sentences. In the case of Sentence 1 weight tokens are identified as common nouns (i.e. *competence* and *kindness*), while in Sentence 2 they are adverbs denoting an exclamatory or interjective function (i.e. because), and redundant and entailed adverbs (i.e. *definitely*, *respectively* [51]).

In the latter situation, therefore, a mechanism can be detected that is intended to explain the rest in more detail, detailing it, and therefore closer to a truthful situation that is always more precisely described than misleading descriptions that are harbingers of ambiguity as already pointed out extensively in the literature. In conclusion, this case history could be an interesting cue for further future investigation and employ such a tendency as a ''tracer'' of model behavior through the bias of precise grammatical elements.

VI. CONCLUSION

The work proposed in this paper contributes to enriching the landscape of application areas of artificial intelligence techniques, with particular reference to the field of NLP. In detail, the in-depth techniques were tested through a new data set created ad hoc for the detection of deceptive reviews, helping to increase the availability of data specific to both the field to which the reviews refer, i.e. cultural heritage, and a low-resource language such as Italian. In addition, some performance baselines were formed through modern language models based on deep learning techniques such as BERT and ELECTRA, adding experiments related to the possibility of exploiting information content related to sentiment polarity to detect the genuineness of reviews, obtaining promising results in these directions.

One of the major limitations of the proposed approach is related to the size of the data set employed, so its expansion is posed as a goal for future work, with the hope of finding more Italian-language data sets in the literature for the cultural heritage sector. An additional reason for future investigation is related to the possibility of extending the baselines and adding further comparisons with other recent language models available in Italian.

From a general point of view, certainly the sentiment classification could benefit from a more granular approach, using a more accurate classification scheme as in [52], as well as further exploration could concern the use of topic modelling to identify themes and areas of user interest within reviews. Finally, the use of tools such as ChatGPT for future research and e.g. the generation of fake reviews could provide further yardsticks of comparison for the proposed dataset.

REFERENCES

[1] A. Ganzaroli, I. De Noni, and P. van Baalen, ''Vicious advice: Analyzing the impact of TripAdvisor on the quality of restaurants as part of the cultural heritage of Venice,'' *Tourism Manage.*, vol. 61, pp. 501–510, Aug. 2017.

[2] N. D. Rodríguez and G. Pisoni, ''Accessible cultural heritage through explainable artificial intelligence,'' in *Proc. 28th ACM Conf. User Modeling, Adaptation Personalization (UMAP)*, T. Kuflik, I. Torre, R. Burke, and C. Gena, Eds., Genoa, Italy, Jul. 2020, pp. 317–324.

[3] G. Pisoni, N. Díaz-Rodríguez, H. Gijlers, and L. Tonolli, ''Human-centered artificial intelligence for designing accessible cultural heritage,'' *Appl. Sci.*, vol. 11, no. 2, p. 870, Jan. 2021.

[4] A. Mukherjee, V. Venkataraman, B. Liu, and S. N. Glance, ''What yelp fake review filter might be doing?'' in *Proc. 7th Int. Conf. Weblogs Social Media (ICWSM)*, E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff, Eds., Cambridge, MA, USA, Jul. 2013.

[5] M. Ott, Y. Choi, C. Cardie, and T. J. Hancock, ''Finding deceptive opinion spam by any stretch of the imagination,'' in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds., Portland, OR, USA, Jun. 2011, pp. 309–319.

[6] M. Ott, C. Cardie, and T. J. Hancock, ''Negative deceptive opinion spam,'' in *Proc. Hum. Lang. Technol., Conf. North Amer. Chapter Assoc. Comput. Linguistics*, L. Vanderwende, H. Daumé III, and K. Kirchhoff, Eds. Atlanta, GA, USA: Westin Peachtree Plaza Hotel, Jun. 2013, pp. 497–501.

[7] J. Ni, J. Li, and J. J. McAuley, ''Justifying recommendations using distantly-labeled reviews and fine-grained aspects,'' in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Nov. 2019, pp. 188–197.

[8] D. Martens and W. Maalej, ''Towards understanding and detecting fake reviews in app stores,'' *Empirical Softw. Eng.*, vol. 24, no. 6, pp. 3316–3355, Dec. 2019.

[9] B. Aejas, A. Bouras, A. Belhi, and H. Gasmi, ''Named entity recognition for cultural heritage preservation,'' in *Data Analytics for Cultural Heritage*. Cham, Switzerland: Springer, 2021, pp. 249–270.

[10] M. Altaweel, ''The market for heritage: Evidence from eBay using natural language processing,'' *Social Sci. Comput. Rev.*, vol. 39, no. 3, pp. 391–415, Aug. 2019.

[11] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, ''Machine learning for cultural heritage: A survey,'' *Pattern Recognit. Lett.*, vol. 133, pp. 102–108, May 2020.

[12] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, ''Fake reviews detection: A survey,'' *IEEE Access*, vol. 9, pp. 65771–65802, 2021.

[13] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *J. Big Data*, vol. 2, no. 1, p. 23, Dec. 2015.

[14] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. Int. Conf. Web Search Web Data Mining (WSDM)*, Palo Alto, CA, USA, M. Najork, A. Z. Broder, and S. Chakrabarti, Eds., Feb. 2008, pp. 219–230.

[15] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. 22nd Int. Joint Conf. Artif. Intell. (IJCAI)*, T. Walsh, Ed., Barcelona, Spain, Jul. 2011, pp. 2488–2493.

[16] N. Jindal, B. Liu, and E.-P. Lim, "Finding unusual review patterns using unexpected rules," in *Proc. 19th ACM Conf. Inf. Knowl. Manag. (CIKM)*, J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, Eds., Toronto, ON, Canada, Oct. 2010, pp. 1549–1552.

[17] S. Feng, L. Xing, A. Gogar, and Y. Choi, "Distributional footprints of deceptive product reviews," in *Proc. 6th Int. Conf. Weblogs Social Media*, Dublin, Ireland, J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, Eds., Jun. 2012.

[18] G. Wu, D. Greene, B. Smyth, and P. Cunningham, "Distortion as a validation criterion in the identification of suspicious reviews," in *Proc. 3rd Workshop Social Netw. Mining Anal. (SNAKDD)*, Paris, France, C. Lee Giles, P. Mitra, I. Perisic, J. Yen, and H. Zhang, Eds., Jun. 2009, pp. 10–13.

[19] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Proc. 7th Int. Conf. Weblogs Social Media (ICWSM)*, Cambridge, MA, USA, E. Kiciman, N. B. Ellison, B. Hogan, P. Resnick, and I. Soboroff, Eds., Jul. 2013.

[20] J. Li, C. Cardie, and S. Li, "TopicSpam: A topic-model based approach for spam detection," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Sofia, Bulgaria, vol. 2, Aug. 2013, pp. 217–221.

[21] S. Banerjee, A. Y. K. Chua, and J.-J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proc. 9th Int. Conf. Ubiquitous Inf. Manag. Commun. (IMCOM)*, Bali, Indonesia, D. S. Kim, S.-W. Kim, S.-H. Lee, L. Hanzo, and R. Ismail, Eds., Jan. 2015, pp. 88:1–88:7.

[22] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 13079–13097, Apr. 2021.

[23] M. B. Khalifa, Z. Elouedi, and E. Lefèvre, "Evidential spammers and group spammers detection," in *Proc. 21st Int. Conf. Intell. Syst. Design Appl. (ISDA)*, in Lecture Notes in Networks and Systems, vol. 418, A. Abraham, N. Gandhi, T. Hanne, T.-P. Hong, T. N. Rios, and W. Ding, Eds. Cham, Switzerland: Springer, Dec. 2021, pp. 255–265.

[24] G. C. Harris, "Detecting deceptive opinion spam using human computation," in *Proc. 4th Hum. Comput. Workshop HCOMP@AAAI*, Y. Chen, P. G. Ipeirotis, E. Law, L. von Ahn, H. Zhang, Eds., Toronto, ON, Canada, Jul. 2012.

[25] J. Li, M. Ott, C. Cardie, and H. E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Baltimore, MD, USA, vol. 1, Jun. 2014, pp. 1566–1576.

[26] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, Jeju Island, South Korea, vol. 2, Jul. 2012, pp. 171–175.

[27] S. Banerjee and A. Y. K. Chua, "A theoretical framework to identify authentic online reviews," *Online Inf. Rev.*, vol. 38, no. 5, pp. 634–649, Jul. 2014.

[28] C. L. Cagnina and P. Rosso, "Classification of deceptive opinions using a low dimensionality representation," in *Proc. 6th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal. (WASSA@EMNLP)*, A. Balahur, E. Van der Goot, P. Vossen, and A. Montoyo, Eds., Lisbon, Portugal, Sep. 2015, pp. 58–66.

[29] A. Molla, Y. Biadgie, and K.-A. Sohn, "Detecting negative deceptive opinion from tweets," in *Mobile and Wireless Technologies*. Singapore: Springer, Jun. 2017, pp. 329–339.

[30] M. R. Martinez-Torres and S. L. Toral, "A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation," *Tourism Manage.*, vol. 75, pp. 393–403, Dec. 2019.

[31] R. N. Zaeem, C. Li, and K. S. Barber, "On sentiment of online fake news," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, The Hague, The Netherlands, M. Atzmüller, M. Coscia, and R. Missaoui, Eds., Dec. 2020, pp. 760–767.

[32] S. Hegde, R. R. Rai, P. G. Sunitha Hiremath, and S. Gangisetty, "Fake review detection using hybrid ensemble learning," in *Advances in Computing and Network Communications*, S. M. Thampi, E. Gelenbe, M. Atiquzzaman, V. Chaudhary, and K.-C. Li, Eds. Singapore: Springer, 2021, pp. 259–269.

[33] Y. Shang, M. Liu, T. Zhao, and J. Zhou, "T-BERT: A spam review detection model combining group intelligence and personalized sentiment information," in *Proc. 30th Int. Conf. Artif. Neural Netw. (ICANN)*, in Lecture Notes in Computer Science, Bratislava, Slovakia, vol. 12895, I. Farkas, P. Masulli, S. Otte, and S. Wermter, Eds. Cham, Switzerland: Springer, Sep. 2021, pp. 409–421.

[34] P. K. Jain, R. Pamula, and S. Ansari, "A supervised machine learning approach for the credibility assessment of user-generated content," *Wireless Pers. Commun.*, vol. 118, no. 4, pp. 2469–2485, Jun. 2021.

[35] P. K. Jain, R. Pamula, and G. Srivastava, "A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews," *Comput. Sci. Rev.*, vol. 41, Aug. 2021, Art. no. 100413.

[36] P. K. Jain, W. Quamer, V. Saravanan, and R. Pamula, "Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis," *J. Ambient Intell. Humanized Comput.*, Jan. 2022.

[37] P. K. Jain, W. Quamer, and R. Pamula, "Consumer sentiment analysis with aspect fusion and GAN-BERT aided adversarial learning," *Expert Syst.*, vol. 40, no. 4, Feb. 2023, Art. no. e13247.

[38] H. Ahsan, V. Kumar, and C. V. Jawahar, "Multi-label annotation of music," in *Proc. 8th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Kolkata, India, Jan. 2015, pp. 1–5.

[39] J. Lee, W. Seo, J.-H. Park, and D.-W. Kim, "Compact feature subset-based multi-label music categorization for mobile devices," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 4869–4883, Feb. 2019.

[40] R. B. Mangolin, R. M. Pereira, A. S. Britto, C. N. Silla, V. D. Feltrim, D. Bertolini, and Y. M. G. Costa, "A multimodal approach for multi-label movie genre classification," *Multimedia Tools Appl.*, vol. 81, no. 14, pp. 19071–19096, Jun. 2022.

[41] J. Akbar, E. Utami, and A. Yaqin, "Multi-label classification of film genres based on synopsis using support vector machine, logistic regression and Naïve Bayes algorithms," in *Proc. 6th Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Dec. 2022, pp. 250–255.

[42] G. Tsoumakas, I. Katakis, and P. I. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., 2nd ed. Boston, MA, USA: Springer, 2010, pp. 667–685.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, J. Burstein, C. Doran, and T. Solorio, Eds., vol. 1, Jun. 2019, pp. 4171–4186.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.

[45] M. Schuster and K. Nakajima, "Japanese and Korean voice search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 5149–5152.

[46] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with Gaussian error linear units," 2016, *arXiv:1606.08415*.

[47] K. Clark, M.-T. Luong, V. Q. Le, and D. C. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.

[48] J. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, C. A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[49] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017.

[50] E. Kokalj, B. Skrlj, N. Lavrac, S. Pollak, and M. Robnik-Sikonja, "BERT meets Shapley: Extending SHAP explanations to transformer-based classifiers," in *Proc. EACL Hackashop News Media Content Anal. Automated Report Gener. (EACL)*, H. Toivonen and M. Boggia, Eds., Apr. 2021, pp. 16–21.

[51] A. T. Neto, "'Adverbs and functional heads' twenty years later: Cartographic methodology, verb raising and macro/micro-variation," *Linguistic Rev.*, vol. 39, no. 2, pp. 293–331, 2022.
[52] X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in *Proc. 5th Workshop Noisy User-Generated Text, W-NUT@EMNLP*, H. Kong, W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, Eds., Nov. 2019, pp. 34–41.

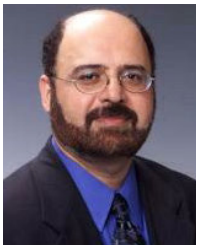
ROSARIO CATELLI is currently a Postdoctoral Research Fellow with the Institute for High Performance Computing and Networking (ICAR), which is part of the National Research Council (CNR), and an Adjunct Professor with the University of Naples Federico II. His research interests include natural language processing, artificial intelligence, and deep learning, also applied to privacy and information security issues.

LUCA BEVILACQUA, photograph and biography not available at the time of publication.

NICOLA MARINIELLO, photograph and biography not available at the time of publication.

VLADIMIRO SCOTTO DI CARLO, photograph and biography not available at the time of publication.

MASSIMO MAGALDI, photograph and biography not available at the time of publication.



HAMIDO FUJITA (Life Senior Member, IEEE) received the master's and Ph.D. degrees in information engineering from Tohoku University, Sendai, Japan, in 1985 and 1988, respectively, the Doctor Honoris Causa degree from Óbuda University, Budapest, Hungary, in 2013, and the Doctor Honoris Causa degree from Timisoara Technical University, Timisoara, Romania, in 2018. He was an Adjunct Professor of computer science and artificial intelligence with Stockholm University, Stockholm, Sweden; the University of Technology Sydney, Ultimo, NSW, Australia; National Taiwan Ocean University, Keelung, Taiwan; and the National Taipei University of Technology. He is currently a Distinguished Professor of artificial intelligence with Iwate Prefectural University, Takizawa, Japan, and a Research Professor with the University of Granada, Granada, Spain; HTECH University, Vietnam; Harbin Engineering University, China; and the Malaysia–Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia. He has supervised Ph.D. students jointly with the University of Laval, Quebec City, QC, Canada; the University of Technology Sydney; Oregon State University, Corvallis, OR, USA; and the University of Paris 1 Pantheon-Sorbonne, Paris, France. He headed a number of projects, including intelligent HCI, a project related to mental cloning for healthcare systems as an intelligent user interface between human users and computers, and SCOPE project on virtual doctor systems for medical applications. He was a recipient of the title of Honorary Professor from Óbuda University, in 2011, and the Honorary Scholar Award from the University of Technology Sydney, in 2012. He was the Editor-in-Chief of *Knowledge-Based Systems* from 2005 to 2020. He is the Emeritus Editor of *Knowledge-Based Systems* and the Editor-in-Chief of *Applied Intelligence* (Springer) and *International Journal of Healthcare Management* (Tayler&Francis). He is also a Highly Cited Researcher in crossfield for the year 2019 and in computer science for the years 2020, 2021, and 2022 from Clarivate Analytics. He is the Chairperson of the i-SOMET Incorporated Association. He has given many keynotes at many prestigious international conferences on intelligent systems and subjective intelligence.

GIUSEPPE DE PIETRO is currently the Director of the Institute for High Performance Computing and Networking (ICAR), National Research Council (CNR), and an Adjunct Professor with the College of Science and Technology, Temple University, Philadelphia, PA, USA. He has been actively involved in many European and national projects, with industrial co-operations. He is the author of more than 200 scientific articles published in international journals and conferences. His current research interests include cognitive computing, clinical decision support systems, and software architectures for e-health. He is a KES International Member. He is also involved in many program committees and journal editorial boards.

MASSIMO ESPOSITO received the M.Sc. degree (cum laude) in computer science engineering from the University of Naples Federico II, in March 2004, and the master's degree (Hons.), named European Master on Critical Networked Systems, and the Ph.D. degree in information technology engineering from the University of Naples, Parthenope, in December 2007 and April 2011, respectively. Since 2012, he has been a Contract Professor of informatics with the Faculty of Engineering, University of Naples Federico II. Since 2020, he has been the Leader of the "Language and Knowledge Engineering" Group, Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). He is currently a Senior Researcher with ICAR-CNR. He has been involved in different national and European projects. He is the author of more than 100 peer-reviewed papers on international journals and conference proceedings. His current research interests include artificial intelligence, natural language processing, knowledge engineering, and conversational systems. He is a member of the editorial board of some international journals. He has been on the program committee of many international conferences and workshops.

Open Access funding provided by 'Consiglio Nazionale delle Ricerche-CARI-CARE-ITALY' within the CRUI CARE Agreement