IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 18, 2020, accepted October 29, 2020, date of publication November 25, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040604

Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection

GIUSEPPE SANSONETTI[1], FABIO GASPARETTI[1], GIUSEPPE D'ANIELLO[2], (Member, IEEE), AND ALESSANDRO MICARELLI[1]

[1]Department of Engineering, Roma Tre University, 00146 Rome, Italy
[2]Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy

Corresponding author: Giuseppe Sansonetti (gsansone@dia.uniroma3.it)

ABSTRACT Since the harmful consequences of the online publication of fake news have emerged clearly, many research groups worldwide have started to work on the design and creation of systems able to detect fake news and entities that share it consciously. Therefore, manifold automatic, manual, and hybrid solutions have been proposed by industry and academia. In this article, we describe a deep investigation of the features that both from an automatic and a human point of view, are more predictive for the identification of social network profiles accountable for spreading fake news in the online environment. To achieve this goal, the features of the monitored users were extracted from *Twitter*, such as social and personal information as well as interaction with content and other users. Subsequently, we performed *(i)* an offline analysis realized through the use of deep learning techniques and *(ii)* an online analysis that involved real users in the classification of reliable/unreliable user profiles. The experimental results, validated from a statistical point of view, show which information best enables machines and humans to detect malicious users. We hope that our research work will provide useful insights for realizing ever more effective tools to counter misinformation and those who spread it intentionally.

INDEX TERMS Deep neural networks, fake news, machine learning, social media.

I. INTRODUCTION

With the spread of social media platforms and the increase of time spent on them, users inevitably exploit such tools for many needs that in the past were otherwise satisfied. Those needs include looking for and reading news articles of interest [1]. In addition to traditional communication channels, such as press, TV, and radio, social media have thus gained increasing importance in the dissemination of news, to the point that, in 2018, two-thirds of the adult U.S. population stated to get news on social media.[1] If social media, on the one hand, guarantee easy and fast access to news articles, which can be consulted at any time by anyone with an Internet connection, on the other hand, they favor the spread of false and unverified news, owing to the same ease with which anyone can publish content: everyone, in the form of a single user or even an entity page (e.g., associated with an online newspaper), can share a news article in the online environment [2]. But social media users themselves are more than aware of this point. In the same paper, the authors point out that a majority (57%) of social media consumers expect news there to be largely inaccurate. It follows that the problem of determining whether a given news on social media is reliable or not is certainly of great importance, as is also discriminating reliable users from those who publish mostly fake or unverified content.

In this article, we adopt the definition of fake news proposed in [3], that is, *unintentional as well as deliberate spread of misleading or wrong narrative or facts*. Similar definitions have also been employed in previous works on the same topic [4]–[6]. Consequently, as unreliable users we mean *entities that spread fake news often intentionally* [7]. There are entities with strong convictions and relevant technical means that exploit online social networks for their purposes. Those entities range from ordinary individuals wishing to increase their popularity on the Internet to real organizations aimed at influencing public opinion. Research literature [8]–[10] shows how it is possible to target certain groups of users to promote specific opinions and content.

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

[1]https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/(Accessed: November 23, 2020)

The purpose of the research work documented herein has concerned the realization of a model, based on Deep Learning techniques, able to detect fake news and unreliable users through two elements: the *text content* and the *social context* in which the news is available. For ''social context'' we mean all the information concerning both the user who shared the news and the interaction among users regarding it. All this data contributes to creating a representative model of news on the chosen social network, that is, *Twitter*.[2] More specifically, the contributions of this article are as follows:

- Analysis of the features (extracted from public data and metadata about users) in order to verify if, and to what extent, they are predictive of the reliability of social media users;
- Design, realization, and experimental evaluation on a publicly available dataset, of an architecture based on deep learning for the prediction of the class (reliable/unreliable) to be assigned to the user profile;
- Online study on real users to verify if there is a connection between the most predictive features from an automatic point of view and the most predictive ones from a human point of view.

The rest of this article is structured as follows. Section II presents an overview of some works related to the proposed approach. The system architecture is illustrated and discussed in Section III. In Section IV, we provide the experimental results of both an offline analysis using real public data and an online study involving real users. We draw our conclusions and discuss some possible future works in Section V. Finally, in Appendix we report the online questionnaire administered to a sample of real users to evaluate their ability to determine if a user profile on Twitter is reliable or not.

II. RELATED WORKS

In the research literature, there exist several noteworthy systems that exploit different approaches to predict the news reliability degree based on the text content only (see, for instance, [11]–[16]). The social media nature itself, however, provides further elements of analysis for the model construction compared to the news content alone. Those elements include features related to aspects and behaviors of relevant users in the social ecosystem, analyzing them from different perspectives. Currently, some significant approaches to fake news detection have been proposed that take into account features from the social context [4]. Such approaches fall into two categories: *stance-based* and *propagation-based*.

A. STANCE-BASED APPROACHES

The former approaches consider the user's stance, that is, her point of view or attitude, and exploit it related to users who have commented on a news, to assess the news reliability [17]. The user's attitude towards a topic can be explicit or implicit. Explicit attitudes include direct expressions of emotions or opinions, such as the ''like'', ''angry'', ''sad'' reactions provided by *Facebook*[3] for posts. Implicit attitudes can instead be automatically extracted from the posts on the social platform. Identifying the user's position concerning a specific post means determining whether she is in favor, neutral, or against a certain topic, idea, or person subject of the post. Topic modeling methods, such as the Latent Dirichlet Allocation (LDA), can be applied to detect latent opinions in posts [18]. It has also been proposed to generate a bipartite graph of users and posts on Facebook by using the information on the attitude underlying the ''like'' reaction [19] and, therefore, to adopt a semi-probabilistic supervised model for predicting the probability that such Facebook posts may contain fake news.

B. PROPAGATION-BASED APPROACHES

The propagation-based approaches to fake news detection take advantage of the interconnection of relevant posts on the social platform to predict the news reliability. The underlying assumption is that the news reliability is strongly correlated with the post reliability in which the news is shared. To analyze the propagation process, homogeneous or heterogeneous reliability networks can be inferred. A homogeneous reliability network consists of a single type of entity [18], such as a post or event, unlike the heterogeneous reliability network involving multiple different types of entity [20]. A proposed technique envisages using a reliability propagation algorithm based on PageRank, by coding the user's reliability and all the implications of posts in a heterogeneous three-level network [21]. Another technique provides for including different aspects of a news to build a three-level hierarchical network and then employing a graph optimization framework to extract the reliability of those aspects [18]. Finally, the relationships between conflicting points of view (i.e., in disagreement with one another) have been used to build a homogeneous reliability network among posts, thus establishing their trustworthiness [18].

Several promising fake news detection approaches based on Graph Neural Networks (GNNs) [22], [23] have recently been proposed, motivated by the latest developments in the domain [24], [25]. Among others, Hu *et al.* [26] propose a model that expands the classical Graph Convolutional Networks (GCNs) proposed by Kipf and Welling [27] to acquire multiscale information of the neighbors based on a given graph. This model, named Multi-depth Graph Convolutional Networks, explicitly preserves the multi-granularity of information, so enhancing the diversity of representation for each node. This allows the approach to improve classification performance and determine the nature of news more effectively than state-of-the-art approaches. In [28], the authors propose a semi-supervised fake news detection method based on GNNs. This method extracts a vector representation of each news through pre-trained GloVe word embeddings, builds a similarity graph between the articles [29], and classifies using two graph neural net-

[2]https://twitter.com/ (Accessed: November 23, 2020)

[3]https://www.facebook.com/ (Accessed: November 23, 2020)

work models: Graph Convolutional Networks [27] and Attention Graph Neural Network [30]. Some approaches derive from recent studies highlighting that fake news and real news spread online in different ways [31], thus determining propagation patterns that can be exploited to detect fake news. The idea of using propagation models to detect fake news has already been explored in several previous studies [32]–[35], in which various types of models have been considered. Propagation-based approaches provide multiple benefits over the content-based approaches, including language independence and better resilience to adversarial attacks [36], [37], in which skilled news makers accurately craft content for avoiding detection. Some of these approaches also employ GNNs. For example, in [3], the authors propose an automatic fake news detection model based on geometric deep learning, a new class of deep learning algorithms designed to deal with graph-structured data [38]. Such algorithms are a generalization of classical GCNs, which allow for the natural integration of heterogeneous data such as content, user profile and activity, social graph, and news propagation. Han *et al.* [39] propose a propagation-based model for fake news detection, which takes advantage of GNNs to discriminate different propagation models of fake and real news on social networks. To this aim, the authors exploit a GNN algorithm designed specifically for graph classification (i.e., the DiffPool algorithm [40]). Furthermore, the authors propose a method that obtains balanced performance on existing and new data, through techniques from continual learning (Gradient Episodic Memory [41] and Elastic Weight Consolidation [42]) to train GNNs incrementally. This avoids retraining the model on the whole data, as it becomes prohibitive as the data size increases. Lu and Li [43] propose a fake news detection model, called Graph-aware Co-Attention Network, which can predict if a short text tweet is fake or not, given the sequence of its retweeters. This model is also capable of generating reasonable explanations (i.e., unveiling why a tweet is fake) through a dual co-attention mechanism, which captures the possible correlation between the tweet and user propagation/interactions. In [44], the authors present a GCN-based approach for detecting rumors on social media. The rationale of this approach is that both propagation and dispersion are essential features of rumors. Therefore, they propose a bidirectional graph model, termed Bi-Directional Graph Convolutional Networks, to explore both features by working on top-down and bottom-up rumor propagation.

Recently, excellent review articles have been published on the topic of online fake news. Among those, Shu *et al.* [4] present algorithms and metrics employed in the detection task, as well as social and psychological aspects underlying the phenomenon. Kumar and Shah [45] analyze the way fake news spreads on the Internet, its impact in economic and social terms, and some approaches that can provide significant performance in identifying misinformation. Zhang and Ghorbani [46] provide the reader with a comprehensive overview of the different aspects of online fake news (i.e., creator, target, and content) as well as the social context in which it proliferates.

The aspects characterizing this work concern both the main objective, that is, to identify unreliable social network profiles rather than fake news, and the choice of features considered to achieve this goal. Those features are related to the news text content, as well as the social context in which the news is spread. The study of their predictivity of the user profile reliability is performed through both an offline analysis conducted on a real dataset collected from Twitter, and an online analysis involving real users. To the best of our knowledge, this is the first study that brings all these characteristics together.

III. SYSTEM ARCHITECTURE

The proposed system performs a double analysis: the prediction of the news reliability and the prediction of the user profile reliability on the social networks. A *reliable* profile is defined in terms of the ratio between the number of real stories deliberately posted and the number of all shared stories. In our evaluation, we took advantage of a popular fact-checker website (see Sec. IV-A) to build up a significant dataset of both reliable and unreliable social network profiles and news. Starting from this dataset, two different processes have been developed. First, we extract features from the news content to create a dataset that can be given in input to various types of classifiers, thus obtaining a result in terms of news classification (fake or real). Then, all the features related to the social context in which the news has spread are extracted and, subsequently, exploited for offline and online analysis. For *offline analysis* we mean an analysis carried out on a dataset through classifiers, having already labeled data, whereas for *online analysis* we mean an analysis performed involving real users, who are given data to be evaluated. The offline analysis, therefore, concerns the creation of a specific dataset containing all the social features of interest, and the training of classifiers through this dataset. The result is a prediction of the user's reliability. Differently, the online analysis is carried out through a questionnaire submitted to 112 real users, so asking them to assess the reliability of a Twitter profile based on the characteristics of the profile itself and the published content on it.

A. NEWS CLASSIFICATION

In order to perform the classification of the news textual content, we used a neural network based on a deep learning architecture (see Fig. 1). It is obtained by combining the properties of a long short-term memory (LSTM) neural network with the properties of a convolutional neural network (CNN). We chose to employ a hybrid approach with these two structures for several reasons including:

- The LSTM network can learn long-term dependencies and is effective for textual data since each word is related to the previous and the following one;
- The LSTM layer allows the model to focus on certain parts in a sequence and to ignore the words unnecessary within the text;

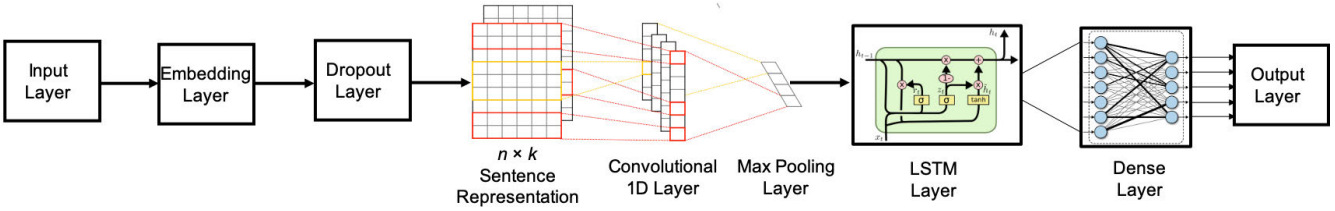


FIGURE 1. Deep neural network architecture employed for news classification.

- The convolution layer is effective in deriving features from a fixed-length segment of the overall input data.

Before the actual training of the network, a tokenization process is applied to the input text, namely, a vocabulary containing the words of the text is built. Inspired by [47], we chose the 30,000 most frequent of them as most significant. In addition to the input and output layers, the network is made up as follows:

- *Embedding layer:* this layer expands each input token into a larger vector, allowing the network to represent words significantly. The first argument was given the value 30,000, which represents the size of our vocabulary, based on the number of words given in input to the previous tokenizer. The second argument used is a value of 128, which means that each token can be expanded into a vector of this dimensionality. The last argument provided is the input length argument set to a value of 1000, which denotes the length of each sequence given in input. Actually, tweets can contain 280 characters at most. However, we include in the analysis also the text of the page linked by the tweet, if present. For that reason, the text input is set to 1000, in order to include the initial content of the linked page, which is considered the most relevant for the classification task;
- *Dropout layer:* this layer works as a regularization technique to reduce the model complexity and prevent overfitting. Dropout can assume a value between 0 and 1 and denotes the fraction of units to be released for the input linear transformation. In our case, we obtained the best results by empirically setting this value to 0.2;
- *Conv1D layer:* a CNN layer extracts features from sequences data and maps the internal features of the sequence. A 1D CNN is effective in deriving features from fixed-length segments, especially where it is not so important where the feature is located in the segment. The Conv1D layer includes 64 filter maps, with a kernel size set to 3. The rectified linear unit (ReLU) is the activation function.
- *MaxPooling1D layer:* this layer allows for the reduction of the input size, thus reducing the number of model parameters (down-sampling) and generalizing the result. It takes the size of the maximum pooling window (set to 4) as hyperparameter, which specifies the selection of the maximum element for each spatial portion identified within the previously generated ReLU map;
- *LSTM layer:* this layer allows the already learned information to persist in the model. It takes the size of the input space (set to 128) as hyperparameter, which denotes the size of the word vector previously defined in the embedding layer;
- *Dense layer:* The result of the CNN pipeline feeds into a 1-layer fully connected neural network structure that drives the final classification decision. The input of this dense layer is flattened into a single vector of values, each representing a probability that a certain feature identified by the deep network identifies a fake news or not. The output layer has one single unit since the system predicts a single value associated with the reliability degree of a news item, represented by values between 0 and 1. The only hyperparameter of this layer is the activation function, which is set to a sigmoid function.

After creating the layers, we moved on to the second step, that is, running the model. To this aim, we chose the hyperparameters as follows:

- *Loss:* we chose binary cross-entropy as loss function because we have only two output classes (0 for fake prediction and 1 for real prediction);
- *Optimizer:* we chose Adam [48] as optimization algorithm, that is, an extension of the stochastic gradient descent procedure. The additional hyperparameters for the Adam optimizer are as follows: the learning rate $\alpha = 0.001$, the exponential decay rate for the first moment estimates $\beta_1 = 0.9$, the exponential decay rate for the second-moment $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$ to prevent any division by zero;
- *Metric:* we chose accuracy as a metric because we are interested in assessing how many positive or negative predictions our neural network can correctly predict.

The last step consists of specifying (through a fit function) how to divide the dataset into a training set and a test set to perform the accuracy evaluation. We chose a partition of 80% (training set) and 20% (test set) because this configuration gave us significant values of prediction accuracy.

B. USER PROFILE CLASSIFICATION

In order to classify the user-related social features, we proposed a system based on deep neural networks. Its performance was then experimentally compared with that of three classifiers used as baselines.

1) NEURAL NETWORK CLASSIFIER (NN)

The operation before the construction of the real model is the normalization. The nature of the data derived from the social features, as we will see later, does not allow for effective training of the classifier. Thereafter, layers are added to the model in an alternation of dense and dropout layers. The number N_h of neurons taken as hyperparameter in the dense layer is 20, chosen based on the following formula [49]:

$$N_h = \frac{N_s}{(\alpha(N_i + N_o))} \tag{1}$$

where N_i is the number of input neurons, N_o the number of output neurons, N_s the number of records in the training dataset, and α an arbitrary factor generally chosen between 2 and 0. In this case, we chose 2, which is considered to be the best value for counteracting overfitting [49]. The network layers are arranged as follows:

1) The first dense layer is a 1-layer fully connected neural network structure that drives the final classification decision. It takes as input the number of neurons (20) and the ReLU activation function. The output layer has the same number of neurons;
2) The second layer is a dropout layer;
3) The third layer is a fully connected layer consisting of 20 neurons, with the ReLU activation function. The output layer has the same number of neurons;
4) The fourth layer is a dropout layer;
5) The last layer is a fully connected layer consisting of 20 neurons, with the sigmoid activation function.

As for the execution of the model above, the hyperparameters are the same as those of the neural network created for news classification.

2) BASELINE CLASSIFIERS

To fully evaluate the potential of the proposed system, we performed a comparative evaluation with three traditional classifiers.

a: LINEAR SUPPORT VECTOR MACHINE CLASSIFIER (SVC)

This classifier is based on a classic linear support vector machine (SVM), which has proven to work well for text classification [50]. The implemented model aimed to find the hyperplanes that best separate training data transformed into coordinates in a high-dimensional space. In particular, a linear kernel has been chosen for its simplicity and high performance, and because the number of features is relatively large, which makes it too difficult to find the optimal hyperparameters when using nonlinear SVMs [51].

b: SUPPORT VECTOR MACHINE CLASSIFIER OPTIMIZED BY STOCHASTIC GRADIENT DESCENT (SVM-SGD)

This implemented model is based on a support vector machine classifier optimized by stochastic gradient descent (SVM-SGD). The hyperparameters we used in this model were:

- *Loss:* defines the loss function l of the model, which for a classification y and an expected output $t = \pm 1$ is defined as follows:

$$l(y) = \max(0, 1 - t \cdot y); \tag{2}$$

- *Learning rate:* learning rate η set as follows:

$$\eta = \frac{1}{\alpha(t + t_0)} \tag{3}$$

  where t_0 is chosen by a default heuristic and α is a non-negative constant that controls the regularization strength (we set $\alpha = 0.0001$);
- *Penalty:* hyperparameter that specifies the regularization term to use, we added a regularization term to the cost function equal to half the square of the L2 norm of the weight vector;
- *Max_iter:* it consists of the number of epochs, that is, the number of complete passes through the training dataset, and a value equal to 5 has been chosen for optimality.

c: K-NEAREST-NEIGHBOR CLASSIFIER (KNN)

This classifier relies on a traditional k-nearest neighbors (kNN) algorithm using the Euclidean distance on the input vectors. An input vector is classified employing the distance from its neighbors, with the vector being assigned to the class most common among its k nearest neighbors. In our experiments, we empirically set k to five. One advantage of this approach lies in its computational efficiency, especially if the same input vector is processed multiple times so that part of the calculus of the distances between vectors can be pre-computed.

Actually, we tested other classifiers (e.g., Logistic Regression and Random Decision Forest) as well as other optimizers, but the difference in terms of performance was not significant. Therefore, we reported only the results of the linear support vector machine classifier (SVC), the support vector machine classifier optimized by stochastic gradient descent (SVM-SGD), and the k-nearest neighbor (KNN).

IV. EXPERIMENTAL EVALUATION

In this section, we report and discuss the experimental results of an offline analysis on a real-world dataset and an online analysis with real users. The values of the hyperparameters reported in Section III-A and considered in the first analysis were obtained through a traditional grid exhaustive grid search technique for the generation of candidates and selection of the best combination.

A. DATASET

Since our goal is to implement methods for detecting fake news and unreliable users, we collected a dataset containing both elements: user profiles and shared news. We took

advantage of *PolitiFact.com*[4] fact-checker focused on the accuracy of statements and news about U.S. politics. Furthermore, we made use of Twitter APIs to collect a set of tweets from September 21, 2019, to November 29, 2019. Each tweet has a reference to a news indexed by the *PolitiFact.com* website, which is classified as real or fake. From this initial dataset, we also extracted content relevant for representing news and the social context of the user sharing it. The final dataset contains 4,022 user profiles, half of which mainly publish fake news, and the other half mainly publish real news. For the sake of completeness, we report here all the dataset statistics:

- 568,315 tweets that reference news indexed on *PolitiFact.com*;
- 62,367 distinct news referenced by tweets and classified as follows:
  - 34,429 fake news;
  - 29,938 verified news;
- 4,022 user profiles classified as follows:
  - 2,013 user profiles who publish mostly fake news;
  - 2,008 user profiles who publish mostly real news.

B. FEATURES

To represent users, many features were selected, extracted, and stored in a matrix that, together with other data, acted as a training dataset for classification models. The features that define the user's social context are the following ones:

- *Screen_name_length:* length of the user's screen name, which is the unique identifier of the profile;
- *Digits_screen_name:* number of numeric characters in the screen name;
- *User_name_length:* length of the user's name, just as she chose it (it cannot be unique);
- *Bio_length:* length of the biography, that is, the short description that the user can add to her profile. If it is not present, the value 0 is entered;
- *Followings:* number of profiles followed by the user;
- *Followers:* number of profiles that follow the user;
- *Favorites:* number of tweets to which the user has put "like" since she signed in;
- *Statuses:* total number of tweets (including retweets) posted by the user;
- *User_listed:* number of lists the user is a member of;
- *Account_age:* number of days the account exists;
- *Statuses/day:* number of tweets that the user posts per day on average.

Furthermore, we consider an additional feature related to the user's sentiment:

- *Sentiment_score:* value between -1 (negative) and $+1$ (positive) that expresses the overall emotion that transpires from the text.

Several additional and potential features have not been considered in the analysis. Whereas our Machine Learning based approach is suitable to be easily extended by considering any characteristics that can be measured in the social network analysis, the experiments include the subset of features to which a consistent measure that is not influenced by any sort of subjectivity can be assigned. For instance, racist, sexist and aggressive language might be linked to content that aims at spreading misinformation, but the scarcity of Natural Language Processing tools and standard measures to identify and measure those dimensions do not allow us to have significant and reliable input for the classification. For that reason, we decided to select the features that are commonly considered in the literature (see, for instance, [52]) for representing tweets and related text content in social networks.

The selected features were stored in a $m \times s$ matrix, with m the number of users and s the number of social features. A second matrix represents the news published by users: for each user, we considered 1 if she has published the news, 0 otherwise. In this way, however, the resulting matrix would be too large and sparse. The news extracted from the starting dataset are 62,367. For this reason, we decided to cluster news through the k-means algorithm, thus obtaining a final matrix containing a reduced number of elements. To identify the optimal number of clusters, we performed the Elbow Method, which varies the number of clusters in a range [1,30], obtaining an optimal value of 24 clusters. The result is a matrix $m \times c$, with m the number of users and c the number of clusters set to the optimal value, in which the number of news that each user has read belongs to each cluster. First, the correlation between the various features was calculated through the creation of a correlation matrix based on the Pearson's correlation coefficient. It should be noted in Table 1 how some features are unrelated to each other or little unrelated, whereas others show high correlation values, such as (listed with the decreasing correlation value):

- *Statuses* $\leftrightarrow$ *Statuses_day* (these two features are clearly related to each other);
- *Followers* $\leftrightarrow$ *User_listed*;
- *Followings* $\leftrightarrow$ *Bio_length*;
- *Statuses* $\leftrightarrow$ *Followers*;
- *Screen_name_length* $\leftrightarrow$ *User_name_length*;
- *Sentiment_score* $\leftrightarrow$ *Bio_length*.

Then, the predictivity of every single feature was also assessed with the 0/1 label (user's reliability/non-reliability) through the Pearson's correlation coefficient. As shown in Figure 2, the most predictive features are the number of followings, followers, and favorites, the number of tweets posted on average per day, and the sentiment score. The other features show a low correlation value.

d: VARIANT WITH INFLUENCER OF THE SOCIAL CONTEXT DATASET

The dataset for the social context analysis, as defined, does not take into account the interaction between users, or retweets. Therefore, we decided to modify the obtained dataset to create a second one for comparative analysis.

[4]https://www.politifact.com/ (Accessed: November 23, 2020)

TABLE 1. Correlation matrix of the user features.

	Screen Name Length	*Digits Screen Name*	*User Name Length*	*Bio Length*	*Followings*	*Followers*	*Favorites*	*Statuses*	*User Listed*	*Account Age*	*Statuses/ Day*	*Sentiment Score*
Screen Name Length	1	0.012	0.406	0.042	-0.003	-0.071	0.015	-0.077	-0.034	-0.198	-0.011	0.062
Digits Screen Name	0.012	1	-0.163	-0.117	0.006	-0.049	0.028	0.005	-0.026	-0.139	0.025	0.001
User Name Length	0.406	-0.163	1	0.131	-0.005	0.016	-0.034	0.001	-0.017	0.055	-0.002	0.014
Bio Length	0.042	-0.117	0.131	1	0.464	0.015	0.029	0.018	0.027	0.097	-0.016	0.399
Followings	-0.003	0.006	-0.005	0.464	1	0.004	0.081	0.063	0.014	0.109	0.032	0.012
Followers	-0.071	-0.049	0.016	0.015	0.004	1	0.008	0.441	0.815	0.076	0.012	-0.004
Favorites	0.015	0.028	-0.034	0.029	0.081	0.008	1	0.137	-0.004	0.017	0.135	-0.117
Statuses	-0.077	0.005	0.001	0.018	0.063	0.441	0.137	1	0.062	0.089	0.919	0.013
User Listed	-0.034	-0.026	-0.017	0.027	0.014	0.815	-0.004	0.062	1	0.099	-0.018	-0.007
Account Age	-0.198	-0.139	0.055	0.097	0.109	0.076	0.017	0.089	0.099	1	-0.084	-0.009
Statuses/ Day	-0.011	0.025	-0.002	-0.016	0.032	0.012	0.135	0.919	-0.018	-0.084	1	0.025
Sentiment Score	0.062	0.001	0.014	0.399	0.012	-0.004	-0.117	0.013	-0.007	-0.009	0.025	1

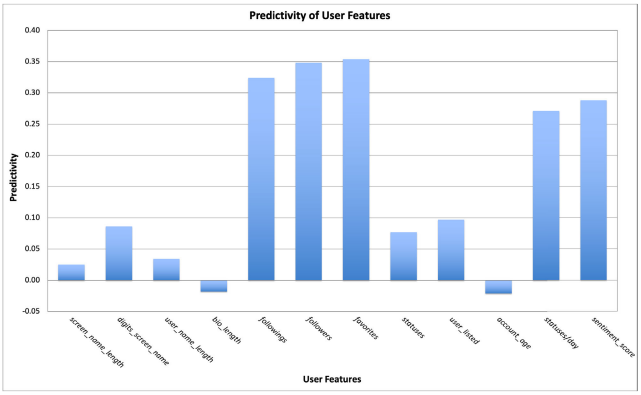


FIGURE 2. Predictivity of the user features for the 0/1 user's reliability class.

In particular, we focused on the values for clusters that contain similar news and, for each pair of users (i, j), if i retweeted j then it means that i was somehow influenced by j. Hence, assuming that the user j has influenced the user i, the vector A_i related to the user i

$$A_i = [v_1, v_2, \ldots, v_n] \tag{4}$$

with $n = 24$ number of clusters, becomes

$$A_i = [v_1 * (\alpha k_1), v_2 * (\alpha k_2), \ldots, v_n * (\alpha k_n)] \tag{5}$$

where $k_1, k_2, \ldots, k_n$ are the values for the n clusters of the user j, and α is a multiplicative factor set to 0.1, a value determined through a mini-batch gradient descent algorithm. A user i can be influenced by multiple users, and in this case, her corresponding A_i vector will be given by the sum of all user contributions she has retweeted.

C. EXPERIMENTAL RESULTS OF THE OFFLINE ANALYSIS

This section first shows the results obtained from the training and validation of the models previously described on the news dataset, then reports the results achieved on the social dataset.

TABLE 2. Results of the four classifiers on the news content dataset.

	Training/Test		**Cross-validation**
	Accuracy	**Loss**	**Accuracy**
NN	91.47%	21.32%	92.89%
SVC	90.02%	17.79%	90.47%
SVM-SGD	89.26%	18.77%	88.51%
KNN	81.54%	28.41%	81.86%

1) RESULTS ON THE NEWS DATASET

News content in the collected dataset has been extracted and used for training the four classifiers. A traditional 80/20 training/test split of the dataset has been considered with 5-fold cross-validation. Table 2 shows the results obtained by each classifier with both validation methods. The best model is the neural network (NN), which shows an accuracy value of 91.47% for the first method, and 92.89% for the second one, despite the loss value is not extremely low. Please note that the loss value that appears in Table 2, as well as in the following ones, refers to the binary cross-entropy of the news classification task (see Sect. III-A). Also, the linear support vector machine classifier (SVC) reported remarkable values: it returns an accuracy value lower than that of the neural network, but a lower loss value. Below, we can find the results of the support vector machine classifier optimized by stochastic gradient descent (SVM-SGD), which provides values slightly lower than the SVC. Finally, the classifier based on the k-nearest-neighbor (KNN) algorithm reported quite satisfactory results even if about 10 percentage points lower than those of the other models.

2) RESULTS ON THE SOCIAL CONTEXT DATASET

The dataset consisting of user information contains categories of values with strongly different ranges. For this reason,

TABLE 3. Results of the NN classifier on the complete social dataset.

	Training/Validate/Test		Cross-validation	
	Accuracy	Loss	Accuracy	Loss
10 Epochs	90.61%	44.31%	91.65%	24.38%
50 Epochs	91.08%	35.82%	92.53%	20.06%
100 Epochs	91.46%	24.47%	93.28%	18.36%

TABLE 4. Results of the NN classifier on the submatrix of the social dataset containing the social features only.

	Training/Validate/Test		Cross-validation	
	Accuracy	Loss	Accuracy	Loss
10 Epochs	80.25%	46.28%	80.32%	44.07%
50 Epochs	81.81%	39.47%	82.05%	38.44%
100 Epochs	82.94%	35.61%	83.08%	33.79%

TABLE 5. Results of the NN classifier on the submatrix of the social dataset containing the news clusters only.

	Training/Validate/Test		Cross-validation	
	Accuracy	Loss	Accuracy	Loss
10 Epochs	88.66%	44.95%	89.18%	29.18%
50 Epochs	90.12%	39.29%	91.95%	26.86%
100 Epochs	90.51%	37.63%	93.04%	25.12%

it was necessary to carry out a normalization in the [0, 1] interval.

a: RUNNING THE NN CLASSIFIER

The tables shown in this section report the results of the first training of the neural network with the complete matrix (see Tab. 3), then with a submatrix formed only by the social features related to the users (see Tab. 4), and finally with a submatrix formed only by clusters of news read/published by users (see Tab. 5). The best results were achieved using the complete dataset. Table 5 shows that also using the submatrix for news clusters, we obtained significant results for both validation methods. As for the submatrix containing the social features, as shown in Table 4, the results were lower than the other two analyses. This can be explained by the fact that the values related to those features, even though normalized, are still more varied than those in the clusters, which vary between 0 and a maximum number of few thousand. Moreover, not all features turn out to be enough predictive of the user's reliability. Nevertheless, it is, however, preferable to also consider the values related to social features in the learning phase since, even if only slightly, they improve the accuracy of the model in addition to the values of the clusters.

b: RUNNING THE SVC, SVM-SGD, AND KNN CLASSIFIERS

For all three baseline classifiers, the best results were obtained on the complete dataset rather than the submatrix of it. The submatrix containing only the features related to the user allowed us to obtain percentage scores, in terms of accuracy, ranging between 70.15% of the KNN classifier (see Tab. 8) and 79.04% of the SVC classifier (see Tab. 6) for the first validation method, and between 70% and 80% for cross-validation. Such not very high scores can be partly explained by the low correlation of some features with the predictive label of the user's credibility, and partly by the fact that the values of those features, even if normalized, are more different between them than those in the clusters. As regards the news clusters, for all three classifiers we indeed achieved satisfactory values, which are close to those obtained on the complete dataset. Among the three classifiers, the best one is based on the linear support vector machine, as occurred for the news content analysis. The worst one is once again the classifier based on the k-NN algorithm.

3) RESULTS ON A VARIANT WITH INFLUENCER OF THE SOCIAL CONTEXT DATASET

We expected better performance on this dataset than the original one from which it was obtained because it contains further information on the interaction between users in the social network. From Table 9, we can see that the results of the deep neural network classifier are remarkable, even better than on the complete dataset without the user's influence (see Tab. 3). In fact, for 100 epochs the accuracy value goes from 91.46% on the first dataset to 92.98% on the second one. The loss value improves as well. The cross-validation shows similar results. As for the three baseline classifiers, if we consider the 80/20 method, also in this case the results exceed the already noteworthy ones obtained on the dataset without interaction between users. If we consider the cross-validation method, the accuracy value remains almost unchanged (i.e., 91%) for the SVC classifier, whereas improves from 89% to 91% for the SVM-SGD classifier. For the KNN classifier, the accuracy value significantly improves, going from 79% to 84% with cross-validation and from 78.61% to 83.13% with the 80/20 method. Generally speaking, this dataset, therefore, enabled us to achieve more accurate predictions with a lower loss value than the original one, thus revealing that the features related to the interaction among users in a social environment can be effective for predicting the user's reliability.

D. EXPERIMENTAL RESULTS OF THE ONLINE ANALYSIS

To perform the online analysis concerning the social context in which the news is placed, an online questionnaire was developed and submitted to real users asking them to assess the reliability of some Twitter profiles. The profiles proposed to testers were related to real people or pages (e.g., newspapers) and selected based on the rate of fake or real news they shared while the dataset collection took place. More specifically, for each user, five highly reliable and five highly unreliable profiles were randomly chosen.

1) QUESTIONNAIRE STRUCTURE

The structure of the questionnaire we used to carry out the online analysis is shown in Appendix. More specifically, we first asked users to provide some personal information for statistical purposes, then to give their opinion on ten selected Twitter profiles. Testers were prompted to express their opinion on the social profile reliability in a five-point Likert scale corresponding to the following answers:

TABLE 6. Results of the SVC classifier on the social context dataset.

	Training/Test						Cross-validation
	Precision	Recall	F1-Score	Accuracy	Average Precision	Loss	Accuracy
Complete Dataset	90%	89%	89%	89.72%	87.22%	12.67%	91%
Only Features	78%	77%	78%	79.04%	78.12%	29.98%	80%
Only Clusters	89%	88%	88%	88.96%	86.11%	11.51%	90%

TABLE 7. Results of the SVM-SGD classifier on the social context dataset.

	Training/Test						Cross-validation
	Precision	Recall	F1-Score	Accuracy	Average Precision	Loss	Accuracy
Complete Dataset	85%	86%	86%	86.07%	83.14%	13.39%	89%
Only Features	76%	75%	76%	77.94%	80.01%	25.68%	78%
Only Clusters	81%	80%	81%	80.76%	78.06%	18.55%	87%

TABLE 8. Results of the KNN classifier on the social context dataset.

	Training/Test						Cross-validation
	Precision	Recall	F1-Score	Accuracy	Average Precision	Loss	Accuracy
Complete Dataset	82%	79%	80%	78.61%	78.02%	21.44%	79%
Only Features	70%	71%	70%	70.15%	69.55%	29.03%	70%
Only Clusters	80%	78%	79%	76.43%	75.04%	20.24%	75%

TABLE 9. Results of the NN classifier on a variant with influencer of the social context dataset.

	Training/Validate/Test		Cross-validation	
	Accuracy	Loss	Accuracy	Loss
10 Epochs	90.78%	29.26%	91.79%	20.14%
50 Epochs	92.07%	21.15%	93.05%	16.13%
100 Epochs	92.98%	19.17%	93.74%	11.84%

- *Strongly disagree*;
- *Disagree*;
- *Neither agree nor disagree*;
- *Agree*;
- *Strongly agree*.

Besides, the opinion expressed by the user had to be integrated with at least one reason behind her point of view. More specifically, the user could choose those reasons from a set of possible answers. Such answers were selected based on the features extracted for the social context model, and aimed to verify if the user's response was consistent or not with the real nature of the Twitter profile (i.e., reliable or not), and the predictive values for the features previously inferred. Those possible answers were summarized in the following macro-groups:

- User information (e.g., profile picture, username, profile description, ...);
- Social information (e.g., followers/followings profiles, number of followers/followings/favorites, ...);
- Writing style (e.g., flame, incitement, sarcasm, ...);
- Behavior in sharing posts (e.g., post length, frequency of sharing, ...);
- Reliability of shared content.

2) OBTAINED RESULTS

To carry out the online analysis, we recruited several testers. The questionnaire was entirely filled in by 112 people, mostly male and 18-36 years old, with at least a three-year degree and a medium to high-level knowledge of English. More detailed information on testers can be found in Table 11. Figure 3 shows the results recorded by the testers in evaluating the reliability of Twitter profiles. Moreover, aggregating the answers of agreement and disagreement (see Fig. 4), we can note how the recognition of unreliable profiles was slightly simpler than that of reliable profiles. It should be also noted the high number of the answer "neither agree nor disagree" chosen by testers, which highlights their difficulty in discriminating the unreliable profiles from the reliable ones. Heterogeneous results have been obtained from the answers concerning the more significant elements in the evaluation of the profile (see Fig. 5). More specifically, user information was the most important reason for both categories, especially for recognizing reliable profiles. In the second place, we can find the reliability of the shared content and the social information, determining more for the recognition of the reliable profiles than the unreliable ones. Then, we can note the behavior in sharing content, which presents substantially the same values

TABLE 10. Results of the SVC, SVM-SGD, and KNN classifiers on a variant with influencer of the social context dataset.

	Training/Test						Cross-validation
	Precision	Recall	F1-Score	Accuracy	Average Precision	Loss	Accuracy
SVC	90%	90%	90%	90.13%	87.34%	9.48%	91%
SVM-SGD	90%	90%	90%	90.94%	88.07%	10.91%	91%
KNN	84%	83%	84%	83.13%	81.19%	15.86%	84%

TABLE 11. Demographics of the 112 users involved in the online analysis.

	Item	Percentage
Gender	Female	42.9%
	Male	57.1%
	Prefer not to say	0.0%
Age	18-26	37.5%
	27-36	35.7%
	37-50	11.6%
	Over 50	12.5%
	Prefer not to say	2.7%
Education	No degree	3.6%
	High school diploma	21.4%
	Bachelor's degree	28.6%
	Master's degree	37.5%
	PhD	7.1%
	Prefer not to say	1.8%
Profession	Unemployed	3.6%
	Student	42.0%
	Employee	33.0%
	Self-employed	10.7%
	Homemaker	5.4%
	Retired	3.6%
	Prefer not to say	1.8%
English proficiency	Low	13.4%
	Medium	61.6%
	High	24.1%
	Prefer not to say	0.9%
Do you use social networks?	No	13.4%
	Yes, sometimes	37.5%
	Yes, often	49.1%
	Prefer not to say	0.0%
Do you have a Twitter profile?	No	56.3%
	Yes	43.8%
	Prefer not to say	0.0%

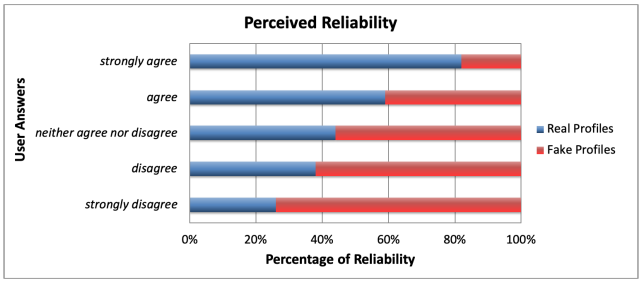


FIGURE 3. Perceived reliability of Twitter profiles with individual data.

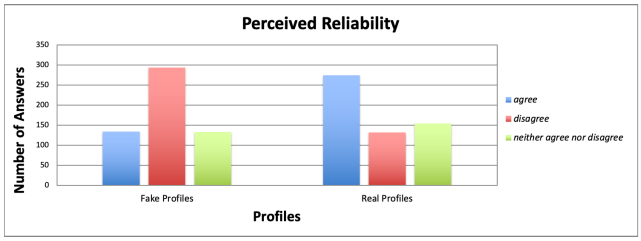


FIGURE 4. Perceived reliability of Twitter profiles with aggregated data.

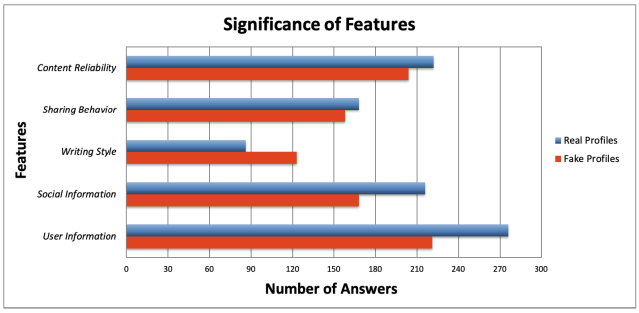


FIGURE 5. Significance of features in the online analysis.

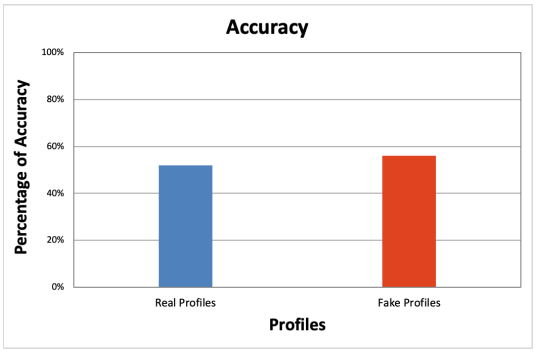


FIGURE 6. Results of the online analysis in terms of accuracy.

for both types of profiles, and finally the writing style, the less voted feature for both real and fake profiles.

The results obtained from the questionnaire regarding the significance of the features partly reflect the predictive values of the features concerning the real or fake label in the offline classification. The most predictive features of the user's reliability were the number of followings, the number of followers, the number of favorites, which are part of the social information; the number of statuses published on average per day, which falls within the behavior of sharing tweets; finally, the sentiment score that had a lower value than the other features, which is reflected in the low number of votes in favor of the writing style. All the experimental results were tested for statistical significance through a two-tailed t-test with a significance level set to $\alpha = 0.01$. We can, therefore, conclude the discussion on the results of the online analysis by pointing out that the collected data shows a non-optimal but still positive recognition of the reliability of the Twitter profiles taken into consideration, compared to the excellent results obtained in the offline analysis. Figure 6 shows that the overall average accuracy obtained in the online analysis was 54%, with a slight predominance in the case of Twitter profiles polarized towards fake news (56%) than profiles polarized towards real news (52%). Furthermore, the importance of the features extracted from the social network for evaluating user profiles has been emphasized, especially as regards user information such as username and description

as well as social information, such as the number of followers/followings.

V. CONCLUSION

The objective of the research work described herein consisted in the study of the features that both from an automatic and a human point of view are more predictive for the identification of social network profiles accountable for spreading fake news. To achieve this goal, we worked on two levels: first, the features related to the news content for the categorization as real or fake were extracted and employed. Subsequently, the focus was on identifying the characteristics of the monitored users, such as social, personal, and interaction information with content and other users, to determine their reliability in sharing news. This last operation made use of two types of analysis: the offline one, performed by training classifiers with labeled data, and the online one, carried out by involving real testers in the evaluation of previously uncategorized data. The results obtained in the experimental evaluations show that for content categorization the objective of discriminating fake from real news has been achieved, with an average accuracy of 90%. Regarding the reliability prediction of the Twitter profiles evaluated in the offline analysis, an accuracy value of about 92% was achieved in the prediction, which is partly reflected in the online analysis (average accuracy of 54%), in which real users were asked to distinguish between real and fake profiles. Moreover, the features recognized as significant for the prediction of the user's reliability partly correspond to the most predictive features related to the real/fake class assigned to each user in the offline analysis.

Although the results obtained can be considered satisfactory, our study suffers from some limitations. Significant among these are the following:

- The dataset on which the experimental evaluations have been set up was collected by querying the *PolitiFact.com* fact-checker service, which is mainly devoted to the U.S. politics. Additional topics such as healthcare and entertainment, are not included in the experiments.
- The online evaluation suffers from skewed demographics: non-U.S. citizens, all English non-native, and mostly in the 18-36 age range. This aspect might affect the outcomes for certain topics.

Nonetheless, we believe that our work will help realize more accurate detection systems of malicious users, by leveraging the features that have proven to be more predictive based on our experimental results.

The future developments that can be applied to this system include, first of all, the use of new datasets of news and users, perhaps collected by a social network of different nature, such as Facebook; the integration with further reliability prediction features such as multimedia content available in news and profiles of users spreading them; the analysis of the social graph including the considered users, based on their interactions such as retweets, mentions, and hashtags; the detection of user communities based on common interests and behaviors, which could set the boundary between users who spread fake content and users who publish real news.

APPENDIX
QUESTIONNAIRE ON TWITTER PROFILES RELIABILITY

In this section, we report the structure of the questionnaire used in the online analysis. We submitted this questionnaire to real users to assess the ability of human beings to distinguish between reliable and unreliable user profiles. Our aim was also to understand which are the most predictive features from a human point of view and if there exists a connection with the most predictive ones from an automatic point of view.

This questionnaire provides ten links to as many Twitter accounts to look over and evaluate on a five-point Likert scale based on various features presented below.

Please, first provide some personal information. We will use them only for statistical purposes.

1. **Gender**
   - ☐ Female
   - ☐ Male
   - ☐ Prefer not to say
2. **Age**
   - ☐ 18-26
   - ☐ 27-36
   - ☐ 37-50
   - ☐ Over 50
   - ☐ Prefer not to say
3. **Education**
   - ☐ No degree
   - ☐ High school degree
   - ☐ Bachelor's degree
   - ☐ Master's degree
   - ☐ PhD
   - ☐ Prefer not to say
4. **Profession**
   - ☐ Unemployed
   - ☐ Student
   - ☐ Employee
   - ☐ Self-employed
   - ☐ Homemaker
   - ☐ Retired
   - ☐ Prefer not to say
5. **English proficiency**
   - ☐ Low
   - ☐ Medium
   - ☐ High
   - ☐ Prefer not to say
6. **Do you use social networks?**
   - ☐ No
   - ☐ Yes, sometimes
   - ☐ Yes, often
   - ☐ Prefer not to say
7. **Do you have a Twitter profile?**
   - ☐ No
   - ☐ Yes
   - ☐ Prefer not to say

Evaluation task (you do not need a Twitter account to take this survey): for each Twitter profile presented below, please indicate whether - in your opinion - the profile is reliable on a scale of five values. Then indicate the reason(s) why you made that choice (you can choose more than one answer to motivate your choice).
(*): mandatory answer

1. **Is this profile reliable? https://twitter.com/…** *
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly agree

   Which elements were decisive in your choice? *
   - ☐ User information (e.g., profile picture, username, profile description, …)
   - ☐ Social information (e.g., followers/following profiles, number of followers/followings/favorites, …)
   - ☐ Writing style (e.g., flame, incitement, sarcasm, …)
   - ☐ Behaviour in sharing posts (e.g., post length, frequency of sharing, …)
   - ☐ Reliability of shared contents
2. **Is this profile reliable? https://twitter.com/…** *
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly agree

   Which elements were decisive in your choice? *
   - ☐ User information (e.g., profile picture, username, profile description, …)
   - ☐ Social information (e.g., followers/following profiles, number of followers/followings/favorites, …)
   - ☐ Writing style (e.g., flame, incitement, sarcasm, …)
   - ☐ Behaviour in sharing posts (e.g., post length, frequency of sharing, …)
   - ☐ Reliability of shared contents
3. **…**
4. **…**
5. **…**
6. **…**
7. **…**
8. **…**
9. **…**
10. **Is this profile reliable? https://twitter.com/…** *
    - ☐ Strongly disagree
    - ☐ Disagree
    - ☐ Neither agree nor disagree
    - ☐ Agree
    - ☐ Strongly agree

    Which elements were decisive in your choice? *
    - ☐ User information (e.g., profile picture, username, profile description, …)
    - ☐ Social information (e.g., followers/following profiles, number of followers/followings/favorites, …)
    - ☐ Writing style (e.g., flame, incitement, sarcasm, …)
    - ☐ Behaviour in sharing posts (e.g., post length, frequency of sharing, …)
    - ☐ Reliability of shared contents

Thank you very much for your cooperation!

ACKNOWLEDGMENT

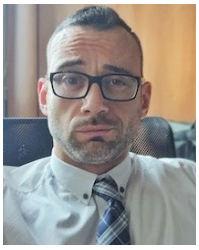
The authors wish to thank the anonymous reviewers for their insightful comments and kind suggestions that allowed them to improve their article. They are also grateful to the colleagues, students, and testers who contributed in some way to their research work.

REFERENCES

[1] K. Shu, S. Wang, D. Lee, and H. Liu, *Mining Disinformation Fake News: Concepts, Methods, Recent Advancements*. Cham, Switzerland: Springer, 2020, pp. 1–19, doi: 10.1007/978-3-030-42699-6_1.
[2] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 5644–5651.
[3] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," 2019, *arXiv:1902.06673*. [Online]. Available: http://arxiv.org/abs/1902.06673
[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, Sep. 2017.
[5] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. ACM Conf. Inf. Knowl. Manage.*, New York, NY, USA, Nov. 2017, pp. 797–806, doi: 10.1145/3132847.3132877.
[6] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "DEFEND: Explainable fake news detection," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Jul. 2019, pp. 395–405, doi: 10.1145/3292500.3330935.
[7] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, p. 109, May 2020, doi: 10.1145/3395046.
[8] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion," *1st Monday*, vol. 21, no. 11, Nov. 2016. [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/7090, doi: 10.5210/fm.v21i11.7090.
[9] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016.
[10] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *J. Econ. Perspect.*, vol. 31, no. 2, pp. 211–236, May 2017.
[11] N. Hassan, C. Li, and M. Tremayne, "Detecting check-worthy factual claims in presidential debates," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2015, pp. 1835–1838.
[12] B. Shi and T. Weninger, "Fact checking in heterogeneous information networks," in *Proc. 25th Int. Conf. Companion World Wide Web (WWW Companion)*, 2016, pp. 101–102.
[13] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2012, pp. 171–175.
[14] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2017, pp. 422–426.
[15] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 231–240.
[16] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as 'False news'" in *Proc. ACM Workshop Multimodal Deception Detection*, New York, NY, USA, 2015, pp. 15–19.

[17] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Trans. Internet Technol.*, vol. 17, no. 3, pp. 1–23, Jul. 2017, doi: 10.1145/3003433.
[18] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2972–2978.
[19] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *CoRR*, vol. abs/1704.07506, 2017. [Online]. Available: http://arxiv.org/abs/1704.07506
[20] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News credibility evaluation on microblog with a hierarchical propagation model," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 230–239.
[21] M. Gupta, P. Zhao, and J. Han, "Evaluating event credibility on Twitter," in *Proc. 12th SIAM Int. Conf. Data Mining (SDM)*. Philadelphia, PA, USA: SIAM, 2012, pp. 153–164.
[22] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.
[23] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014, pp. 1–14, [Online]. Available: http://arxiv.org/abs/1312.6203
[24] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: http://arxiv.org/abs/1812.08434
[25] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 24, 2020, doi: 10.1109/TNNLS.2020.2978386.
[26] G. Hu, Y. Ding, S. Qi, X. Wang, and Q. Liao, "Multi-depth graph convolutional networks for fake news detection," in *Natural Language Processing and Chinese Computing*, J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds. Cham, Switzerland: Springer, 2019, pp. 698–710.
[27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017, pp. 1–14. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl
[28] A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-supervised learning and graph neural networks for fake news detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Vancouver, BC, Canada, Aug. 2019, pp. 568–569, doi: 10.1145/3341161.3342958.
[29] G. B. Guacho, S. Abdali, N. Shah, and E. E. Papalexakis, "Semi-supervised content-based detection of misinformation via tensor embeddings," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 322–325.
[30] K. K. Thekumparampil, C. Wang, S. Oh, and L.-J. Li, "Attention-based graph neural network for semi-supervised learning," 2018, *arXiv:1803.03735*. [Online]. Available: http://arxiv.org/abs/1803.03735
[31] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, May 2018. [Online]. Available: https://science.sciencemag.org/content/359/6380/1146
[32] Y. Liu and Y. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. 32nd AAAI Conf. Artif. Intell., (AAAI), 30th Innov. Appl. Artif. Intell. (IAAI), 8th AAAI Symp. Educ. Adv. Artif. Intell. (EAAI)*, New Orleans, LA, USA, Feb. 2018, pp. 354–361. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16826
[33] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," in *Proc. 14th Int. AAAI Conf. Web Social Media (ICWSM)*, Atlanta, GA, USA, Jun. 2020, pp. 626–637. [Online]. Available: https://aaai.org/ojs/index.php/ICWSM/article/view/7329
[34] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, New York, NY, USA, 2018, pp. 637–645, doi: 10.1145/3159652.3159677.
[35] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD Explor. Newslett.*, vol. 21, no. 2, pp. 48–60, Nov. 2019, doi: 10.1145/3373464.3373473.
[36] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, *arXiv:1412.6572*. [Online]. Available: https://arxiv.org/abs/1412.6572
[37] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014, *arXiv:1312.6199*. [Online]. Available: https://arxiv.org/abs/1312.6199
[38] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.
[39] Y. Han, S. Karunasekera, and C. Leckie, "Graph neural networks with continual learning for fake news detection from social media," 2020, *arXiv:2007.03316*. [Online]. Available: http://arxiv.org/abs/2007.03316
[40] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 4805–4815.
[41] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6470–6479.
[42] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.
[43] Y.-J. Lu and C.-T. Li, "GCAN: graph-aware co-attention networks for explainable fake news detection on social media," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 505–514. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.48/
[44] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," 2020, *arXiv:2001.06362*. [Online]. Available: http://arxiv.org/abs/2001.06362
[45] S. Kumar and N. Shah, "False information on Web and social media: A survey," 2018, *arXiv:1804.08559*. [Online]. Available: http://arxiv.org/abs/1804.08559
[46] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.
[47] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–12.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[49] H. B. Demuth, M. H. Beale, O. De Jess, and M. T. Hagan, *Neural Network Design*, 2nd ed. Stillwater, OK, USA: Martin Hagan, 2014.
[50] Y. Yang, "An evaluation of statistical approaches to text categorization," *Inf. Retr.*, vol. 1, no. 1, pp. 69–90, Apr. 1999, doi: 10.1023/A:1009982220290.
[51] Y. Gao and S. Sun, "An empirical evaluation of linear and nonlinear kernels for text classification using support vector machines," in *Proc. 7th Int. Conf. Fuzzy Syst. Knowl. Discovery*, Aug. 2010, pp. 1502–1505.
[52] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proc. 11th Int. Conf. Web Social Media (ICWSM)*, Montreal, QC, Canada, May 2017, pp. 280–289. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587

GIUSEPPE SANSONETTI received the master's degree *(summa cum laude)* in electronic engineering and the Ph.D. degree in computer science. From 2003 to 2005, he pursued research at the University of California at Santa Barbara (UCSB), Santa Barbara, and Oregon State University (OSU). He has also worked as a Research Fellow with the Italian Institute for Nuclear Physics (INFN) and the National Interuniversity Consortium of Materials Science and Technology (INSTM). He is currently an Assistant Professor with the Artificial Intelligence Laboratory, Department of Engineering, Roma Tre University, where he has been teaching the "Intelligent Systems on the Internet" course for the master's degree in computer science. He has been involved in National and International research projects regarding Internet technologies. His current research interests include user modeling, recommender systems, case-based reasoning, and computer vision. He has been a Reviewer of leading international journals and a member of program committees of relevant international conferences.



FABIO GASPARETTI received the M.S. and Ph.D. degrees in computer science and automation in 2001 and 2005, respectively, doing research on adaptive web search. He was a Visiting Scientist and a Researcher with Nokia Bell Labs, Cambridge, U.K., and Xerox PARC. Since 2017, he has been a member of the A. I. Task Force promoted by Agency for Digital Italy (AgID). He is currently an Assistant Professor with the Artificial Intelligence Laboratory, Department of Engineering, Roma Tre University. His research interests include user modeling and user-adapted interaction; recommender systems and context-awareness; and adaptive focused crawling. He has been a Reviewer of well-established international journals and stably involved in technical program committees of relevant conferences.

GIUSEPPE D'ANIELLO (Member, IEEE) received the degree and the Ph.D. degree in information engineering from the University of Salerno, Italy, in 2013 and 2018, respectively. He is currently an Assistant Professor with the Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno. He has been a Research Scientist with the CORISA (Research Consortium on Agent System), where he was the Project Manager of the Italian Research and Development projects ''MAR.TE.–Meta-managing, re-profiling, and monitoring key sea-land logistics processes." His current research interests include situation awareness, semantic web, computational intelligence, and granular computing. He has coauthored several scientific papers on the aforementioned topics which have been published in international journals and conference proceedings. He is a member of the IEEE SMC Technical Committee on Cognitive Situation Management. He is also a member of the IEEE Computational Intelligence Society, IEEE Systems, Man, and Cybernetics Society, and IEEE Computer Society.

ALESSANDRO MICARELLI is currently a Full Professor of artificial intelligence with the Department of Engineering (DE), Roma Tre University, where he has been teaching the Artificial Intelligence and Machine Learning courses for the ''Laurea Magistrale'' (Master) degree in computer science. He is also in charge of the Artificial Intelligence Laboratory, DE. His long-term research goal is focused on how AI systems can cooperate most effectively with humans, through harnessing their adaptive power to provide human–system interaction that best responds to users' needs, interests, cognitive styles, and background, as well as to the real-world context in which they are operating. Actual projects range from adaptive web-based systems, user modeling, and personalized search to recommender systems, and artificial intelligence in education. He is a founding Member of the Italian Association of Artificial Intelligence. He has helped to organize various international conferences (lately: the General Co-Chair of UMAP 2013 and the Program Co-Chair of ITS 2016) and has peer-reviewed articles for leading international journals. He has participated in various international research projects and was the National Coordinator of a PRIN research project on e-learning funded by the Italian Ministry of Education, University and Scientific Research (MIUR). He is also a Scientific Consultant at MIUR and a member of a Board for the evaluation of Industrial Research proposals.