Implementation of a Multi-Approach Fake News Detector and of a Trust Management Model for News Sources

Article *in* IEEE Transactions on Services Computing · September 2023

DOI: 10.1109/TSC.2023.3311629

CITATIONS
0

READS
48

5 authors, including:

Claudio Marche
Università degli studi di Cagliari
14 PUBLICATIONS **240** CITATIONS
SEE PROFILE

Ilaria Cabiddu
Università degli studi di Cagliari
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Luigi Serreli
Università degli studi di Cagliari
7 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Michele Nitti
Università degli studi di Cagliari
50 PUBLICATIONS **3,558** CITATIONS
SEE PROFILE

Implementation of a multi-approach fake news detector and of a trust management model for news sources

Claudio Marche, *Student Member, IEEE,* Ilaria Cabiddu, Christian Giovanni Castangia, Luigi Serreli, *Student Member, IEEE* and Michele Nitti, *Senior Member, IEEE*

Abstract—Technological development combined with the evolution of the Internet has made it possible to reach an increasing number of people over the years and given them the opportunity to access information published on the network. The growth in the number of fake news generated daily, combined with the simplicity with which it is possible to share them, has created such a large phenomenon that it has become immediately uncontrollable. Furthermore, the quality with which malicious content is made is increasingly high so even professional experts, such as journalists, have difficulty recognizing which news is fake and which is real. This paper aims to implement an architecture that provides a service to final users that assures the reliability of news providers and the quality of news based on innovative tools. The proposed models take advantage of several Machine Learning approaches for fake news detection tasks and take into account well-known attacks on trust. Finally, the implemented architecture is tested with a well-known dataset and shows how the proposed models can effectively identify fake news and isolate malicious sources.

Index Terms—Fake News Detection, Trustworthiness Management, Machine Learning, Prebunking.

1 Introduction

The technological development combined with the evolution of the Internet has made it possible to reach an increasing number of people over the years. The spread of smart devices has allowed users to be able to connect anywhere and anytime to the network: the visible advantages are represented by the opportunities for everyone to access information published on the network, easily increase their cultural background, and make their opinion heard.

This scenario has allowed the birth and creation of new websites that provide large amounts of information, even free of charge, to an ever-growing audience eager to expand their knowledge. However, the simplicity with which it is possible to publish news online has allowed anyone to disseminate news of all kinds so that also the propagation of distortions, alternate realities, and lies has increased. This phenomenon is now known as Fake News, so finding reliable information on the Internet has become problematic.

Fake news is defined as information that is partially or completely false, disseminated intentionally or unintentionally through any means of communication that presents an apparent plausibility and a greater increase in the prejudices that lie with it [1].

Detection algorithms have the crucial task of implementing technical approaches in service provisioning and methodologies to aggregate a variety of information in order to infer the reliability of news the user wishes to interact with. However, debunking is a difficult task and has to overcome several challenges: aside from the size of published fake news to be verified, corrective information can sometimes provoke a so-called "backfire effect" in which respondents more strongly endorse a misperception about a controversial political or scientific issue when their beliefs or predispositions is challenged [2]; finally, debunks do not reach as many people as fake news, and they do not spread nearly as quickly [3].

To this, the goal of this paper is not only to evaluate news items, i.e. to understand if the news is real or fake, but also to develop a prebunking system [4], i.e. the process of debunking lies, fake news or sources before they strike, by evaluating the trustworthiness of the news providers.

Trust is tied to the concept of reputation. Indeed, trust can be gained on both direct and indirect bases, but in large networks such as the Internet, it takes time for a user to collect enough direct experience so an entity has to rely on the perception of other entities, that is the reputation. Through reputation, it is possible to collect, distribute and aggregate feedback about participants' past behaviour and then provide a global perception of an entity. This concept enables newsreaders to rely on the community's reputation to identify trustworthy sources, eliminating the need for a trial-and-error approach. To this, in this paper, we have developed a trust and reputation management model so that it is possible for users to understand which are the news providers that can lead to successful collaboration, i.e. that can provide reliable news.

This paper is part of the project FAKE, developed as a cascade call of the EU's project TruBlo [5]. In particular, this paper provides the following contributions:

1) First, we proposed a detection algorithm that analyses

- *C. Marche, I. Cabiddu, C. G. Castangia, L. Serreli and M. Nitti are with the Department of Electrical and Electronic Engineering (DIEE), University of Cagliari, Italy.*
- *C. Marche and M. Nitti are with National Telecommunication Inter University Consortium, Research Unit of Cagliari, Italy.*

E-mails: C.M. claudio.marche@unica.it, L.S. luigi.serreli@unica.it and M.N. michele.nitti@unica.it.

news' written text and classifies the news as fake or real according to several parameters, which include the writing style, fact-checking, sentiment analysis of the text, and the context of the news.
2) Second, we develop a trust management model for evaluating news sources, which uses novel parameters, namely expertise, relevance, goodwill, and coherence, to defend against malicious behaviours.
3) Finally, we simulate the implemented architecture by using a Kaggle dataset, which contains a total of $20,387$ news from various domains (such as politics and economics) to show the performance of each module of the algorithm and its overall accuracy in identifying fake news.

The rest of this article is organised as follows: Section 2 presents a brief survey on fake news detectors and on trustworthiness algorithms used to classify news providers. In Section 3, we define the system architecture and the reference scenario. Section 4 and 5 present the fake news detector and the trust management algorithm. Furthermore, the system performance is analysed in Section 6, while Section 7 presents an alternative technology for storing and retrieving information related to the news, namely Blockchain, and compares it to a traditional database. Finally, Section 8 draws final remarks.

2 Related Works

2.1 Fake News Detectors

In recent years, there has been a significant focus in the literature on analysing fake news, and numerous works have been proposed to detect them [6]. The growth of social media and the abundance of online information has considerably added complexity to this challenge [7]. When sharing news, people often fail to consider the possibility of fake news and tend to believe only the news that confirms their pre-existing beliefs. This lack of critical thinking leads to a failure to reflect on the reliability and truthfulness of the information they see on social media platforms [8]. Another issue concerns the rapid spread of fake news, which can propagate much faster, deeper, and broader than accurate news, resulting in a significant proportion of the information people encounter daily being false [9], [10]. Furthermore, although fake news is not a new phenomenon, it is rapidly increasing and gaining public attention [11]; the leading cause is that fake news can be created cheaper and faster than traditional news media [12]. In this regard, fake news detection is becoming a critical mechanism that proposes to detect fake content as fast as possible and provide assistance to journalists and fact-checkers [13]. Below, we want to analyse and classify the most important detectors based on their techniques and approaches.

In these terms, two well-known fake news detectors, based on analysing the news features through multiple machine-learning techniques, are illustrated in [14] and [15]. In the first work, the authors make use of different machine-learning approaches as classifiers for fake news considering linguistic and count-based features, such as length and word count. Authors denote how fake news articles usually tend to be shorter, appear more negative, and adopt a more personal disclosing tenor. In the second work, the authors propose a similar solution and demonstrate how leveraging various sources of sentiment, e.g., images and visual media, can be used to improve accuracy. The approach is evaluated using several datasets and similarity techniques. Both works obtain the best results with the Support Vector Machine (SVM) algorithm, which is then used for the fake news detection processes. Moreover, an approach mainly based on SVM is presented in [16], in which the authors propose a fake news detection model based on n-gram analysis, i.e., an approach used in language modelling and natural language processing, combined with a linear SVM (LSVM) classifier. Various sequences of characters or words, namely n-grams, are generated from a training set and compared in order to classify fake from honest news. All the n-grams are then used as input for a machine-learning technique responsible for the final classification.

Two other approaches that mainly focus on machine learning are described in [17] and [18], in which the authors perform the detection through neural network architectures. In the first paper, the authors especially focus on feature extraction, studying the most relevant attributes of text news. They identify different kinds of features: content features, such as the number of words and the frequency of characters, user features, based on the news readers and in particular on the users who have interacted with the news, and, finally, social features, which refer to the social connections of the users. All the features, considering text and news context, are then evaluated and compared using a Convolutional Neural Network (CNN) and a Long short-term memory (LSTM) that provide the news classification. In the second work, the authors propose a classification model for fake news detection based on linguistic features and automatic fact-checking. The model evaluates the news considering linguistic features, such as the number of words and sentences, and then compares them with mainstream verified articles; a deep learning algorithm is trained to learn the common patterns and produce the classification.

Furthermore, two different machine-learning techniques are illustrated in [19] and [20]. In the first approach, the authors propose a fake news detector based on the analysis of term frequency and unique words. After this feature extraction process, a Naive Bayes Classifier (NBC) model is trained and then proceeds to the news classification. The system is evaluated considering the precision and the accuracy of classification. In the second one, the authors analyse the association between fake news and clickbait and how in general, the goal of fake news producers is to profit through clickbait. Clickbait lures users and raises curiosity with flashy ads or designed click links to increase revenues. In these terms, the authors propose a fake news detection model based on context analysis, e.g., collecting URLs commonly used for clickbait and linguistic features, such as the number of capitalised characters or exclamation marks. Moreover, a work based on well-known machine-learning techniques is illustrated in [21]. The authors propose an automatic fake news detection based on BERT and ALBERT models that retrieve the most relevant facts concerning the news claims and verify the level of truth by computing a textual comparison. A series of transformer models are observed and used to compare news and facts retrieved from a manually curated dataset.

TABLE 1: Analysis of existing fake news detectors.

Ref	Quant.	Informal.	Complex.	Divers.	Fact-Checking	Sentim.	Ads
[14]	✓	✓	-	-	-	✓	-
[15]	-	-	✓	-	-	✓	-
[16]	✓	✓	✓	✓	-	-	-
[17]	✓	-	✓	✓	-	-	-
[18]	✓	-	✓	✓	✓	-	-
[19]	✓	-	✓	✓	-	-	-
[20]	✓	-	-	-	-	-	✓
[21]	✓	-	✓	-	✓	-	-
[22]	-	-	✓	-	✓	-	-
[23]	✓	-	✓	✓	-	✓	-
FAKE	✓	✓	✓	✓	✓	✓	✓

The last group of articles focuses on different approaches that do not consider machine learning techniques for fake news detection. In these terms, two works are illustrated in [22] and [23], in which the authors mainly concentrate on sentiment analysis. In the first work, a framework to encourage fact-checked content is proposed, and the authors examine active Twitter users, called guardians, who share validated information in order to correct fake content in online discussions and provide them with a URL-based fact-checking recommendation model to stimulate their engagement and reduce the negative effects of fake news. At first, the proposed model focuses on the detection of the guardians' users, then analyses the textual claims and recommends guardians' fact-checking URLs to the other users. In the second work, the authors propose a model to detect fake news using sentiment analysis as the main feature. The model combines the sentiment related to the text with style features, such as the number and frequency of words and statements, and performs the classification through different machine learning algorithms.

In summary, Table 1 shows a classification of the fake news detection models based on the analysed approaches. The classification is based on three metrics, i.e., text quality, fact-checking, and context analysis. In general, the features needed to measure the news quality can be divided into four main categories [24]: *Quantity*, which considers features such as the number of characters, the number of words and the number of sentences; *Informality*, which takes into account the fact that fake news often contains more mistakes than trustworthy ones, and therefore characteristics such as misspellings and typos are used as indicators of the authenticity of the news; *Complexity*, which is represented by parameters such as the average word length, the words per sentence and the average ratio of punctuation per sentence. In general, the higher the linguistic complexity of a text, the less likely it is a fake. Finally, the last category is depicted by the *Diversity*, which considers the percentage of different terms in the text, the occurrence of the words, and their spatial distance; deceptive texts are perceived to be limited in terms of vocabulary usage and usually make use of several redundant terms. Another classification metric is represented by *Fact-Checking*. Even if the text analysis is largely used for classification, other approaches are used to support it. To verify the correctness of the information, researchers propose systems to check the credibility of the news. The last metric concerns the analysis of the news context. It classifies news through the analysis of the website in which the news is published with features such as the *Sentiment* related to the news content or the presence of advertisements (*Ads*) as well. The solution proposed in this paper takes into consideration all kinds of news topics, and no user information is collected; for this reason, properties related to the news readers are out of the scope of our solution.

2.2 Trustworthiness on News Providers

The issue of trust in news providers has been gaining huge popularity in recent years, and the community is trying to find new approaches to study the news providers' trustworthiness [25]. This is due to the fact that non-verified news media have exploited technological development to spread misleading or fraudulent news [26], [27]. In general, trust in news providers is at historically low levels and new mechanisms to measure their trust are essential to overcome this age of uncertainty [28].

Below, we provide a brief background on the most accepted properties used in literature to evaluate the trust of a news provider. One of these is presented in [29], where authors introduce the concept of trustworthiness in news media as a relationship between a trustee, represented by the user or the actor who trusts, and a trustor, i.e., the news source that provides the news and receives the trust. The news media is then evaluated, and the trust is obtained through 4 dimensions: trust in the selectivity of topics, selectivity of facts, the accuracy of depictions, and journalistic assessments. Another two works that analyze the impact of trust in news media are illustrated in [30] and [31]. The first paper proposes various measures of trust at different levels of analysis. The source is evaluated both generally and in each specific topic, taking into account its fairness and accuracy in distinguishing between facts and opinions. Then, in the same way, the paper judges the author of the news, the journalist, and the media outlet, according to their objectivity and subjectivity. In the second paper, the authors examine the news media trust in terms of credibility in a political topic for five different countries. The authors present a regression model that assesses news providers based on various features, including media attributes such as overall credibility and reading frequency, as well as demographic metrics of the news reader.

In these terms, we propose an automatic algorithm to evaluate the trust of news sources based on the most accredited measures of credibility. It makes use of novel parameters, which consider the expertise, the popularity of the news source, and its past behaviour.

3 Introduction to the Proposed Solution

This section provides a detailed description of the behaviour of the entire system. All the functionalities described here are presented to the final users through a plugin which interacts with the FAKE system.

The FAKE system is comprised of three entities, as shown in Figure 1, which are available after an authentication phase: a fake news detector, a trust management model, and an encrypted database. The fake news detector is responsible for evaluating news. When a user requests

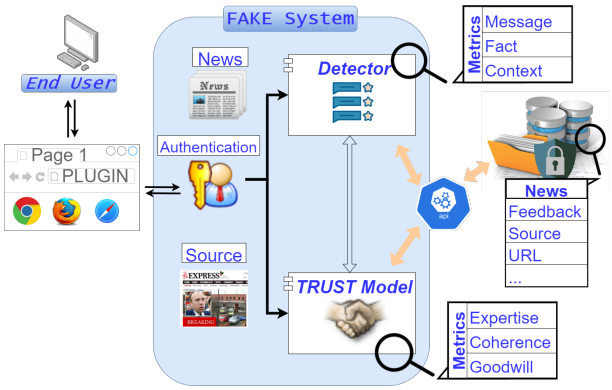


Fig. 1: Overall system architecture.

news, the system initiates the evaluation process to provide feedback on the requested information. To this, the detector extracts important features from each news item and implements several algorithms to provide feedback. The purpose of this entity is then to store the feedback on the database for future uses. The trust management model is designed to evaluate news producers' websites: the model has access to the database in order to retrieve all the information regarding a specific news source, such as feedback or news topics. This information is aggregated to extract novel parameters, which are used to evaluate the reliability of the news source. Finally, the database has the important role of acting as an intermediary between the two previous entities, allowing to save and retrieve the evaluation of the news. To ensure the security of exchanged data, even in the face of potential data leaks, we adopt the Blowfish algorithm to encrypt the results of the evaluation process, which represents the cornerstone of the entire system. We have chosen this algorithm since the encryption is only used internally to the system and there is no need to send feedback data to third parties. However, we only encrypt feedback data, leaving non-sensitive data unencrypted. This approach reduces the computational and time costs associated with encryption, making it an effective means of protecting confidential data [32].

The plugin continuously monitors the web pages visited by the user. The activation of the fake news detector or the trust management model depends on the user's browsing behaviour. To illustrate the system's functionality, we provide two sequence diagrams in Figure 2. Solid arrowheads indicate calls to system entities, while dashed lines represent reply messages. If the plugin detects news (Figure 2a), it triggers the fake news detector. The detector evaluates the reliability of the news: to this, it first checks if the news has already been assessed and if relevant feedback has been stored in the database. In this case, the detector retrieves the feedback value and immediately displays it to the user. However, if no feedback regarding the news is found on the database, the detector starts analysing the news and extracts all the parameters of interest, such as the topics, URL, and source, as well as features related to the text, sentiment, and more. The detector then computes the news feedback based on the model explained in Section 4 and stores the feedback, along with the associated information, in the database: feedback can then be retrieved by any of its parameters, such as the URL of the news, source, topics, timestamp, and so on, so the proposed feedback has multidimensional views. In order to keep our formulas as clean as possible, we will only address the feedback with the minimum notation needed to explain the model, but the reader should keep in mind that the other parameters are only hidden but always available.

Whenever a user is looking for news on a browser or a news producer's website (Figure 2b), the plugin interacts with the trust management system. In order to compute the trustworthiness of the news producers, the trust model retrieves all the needed information from the database and shows the user the risk associated with every news website to provide the user with the best alternatives. The trust of a news source is evaluated by considering all the topics it has covered. This means that a source could be considered reliable for certain topics, but not for others.

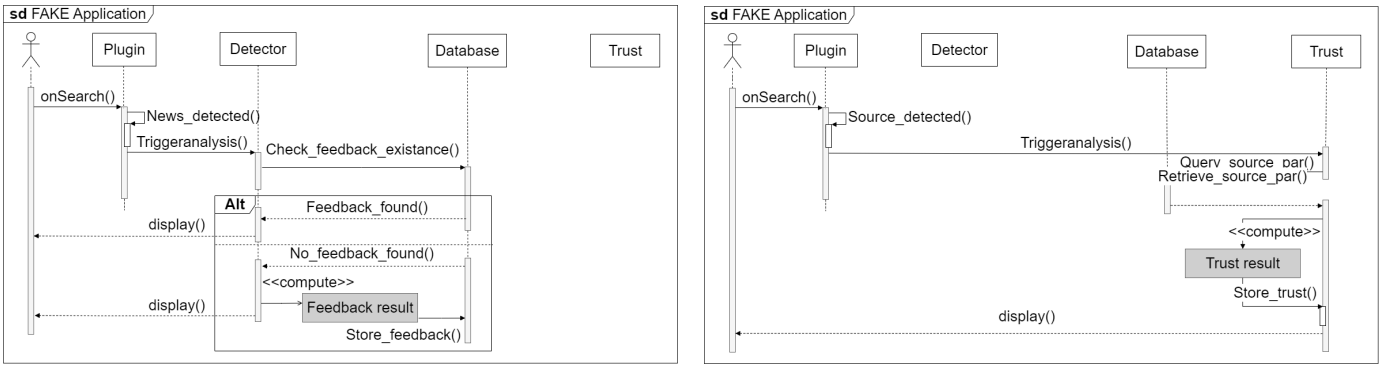
Finally, we note how the collection of personal data takes place only when the user downloads the application, i.e., the plugin, to be integrated into their browser, and it will only be used for contractual purposes. During the browsing phase, no user data is collected, and the only information that the system collects is related to the web page that the user has visited anonymously.

4 Feedback Evaluation Model

According to the presented scenario, in this Section, we propose our feedback evaluation model, which is responsible for assessing the news selected by users. In order to evaluate and assign feedback, the model considers several parameters, which can be classified into three different factors: message-based, fact-based and context-based parameters. The first factor refers to the style of the news and proposes to analyse the text's characteristics based on quantity, informality, complexity and diversity. The second factor considers the examination of facts comparing news claims with a large well-known pre-trained model or with news already evaluated. Finally, the context-based parameters consider the presence of ads on the web page and sentiment analysis, which evaluates the text in terms of sentiment and objectivity as well. Therefore, the feedback f_i related to a news n_i is computed as:

$$f_i = \alpha M_i + \beta F_i + (1 - \alpha - \beta) C_i \tag{1}$$

where M_i represents the contribution of the message-based, while F_i and C_i depict the fact-based and context-based parameters respectively. All these factors, namely M_i, F_i and C_i, are in a range $[-1, 1]$, while the weights $\alpha, \beta \in [0, 1]$ are selected to give more importance to a particular factor. The weights are selected so that their global sum is equal to 1 in order to normalize the feedback value in the continuous range $[-1, 1]$, where the value equals to -1 depicts news generated to harm someone or something, i.e., new articles created to spread disinformation [33], while the unitary feedback value corresponds to reliable information. Among the concept of information and disinformation, values of f_i around zero indicate misinformation, i.e., false information shared without the intention to harm [34]. There exist two zones of uncertainty where the classification is



(a) Fake news detector.

(b) Trust management model.

Fig. 2: Sequence diagrams for the fake news detector and the trust management model.

TABLE 2: Message-based factor features.

Category	Feature	Description
Quantity	Characters	Number of characters in the news text.
	Words	Number of words in the news text.
	Average of words	Average of words per sentence.
	Puntuaction	Average of punctuation with respect to the number of characters.
Informality	Bad words and toxic content	Presence of bad words or abusive language in the message.
	Typos and misspelling	Number of unknown words and check their similarity with others.
Diversity	Redundancy	Occurrence of the words and spatial distance between them.
Complexity	Term frequency	Frequency of words in the whole text.

difficult. These zones represent the transition from disinformation to misinformation and misinformation to information. In order to resolve this uncertainty, we define a threshold TH so that all $f_i : f_i \leq |TH|$ are classified as misinformation.

4.1 Message-based

The Message-based factor represents the first set of parameters related to the analysis of the news through their style which make use of the text quality to distinguish false statements from real ones. At first, the text news is pre-processed in order to clean the text by removing special characters and stopwords, making it ready to feed the text to our model. Then the model proceeds to the feature extraction step so that the writing quality can be measured based on features of quantity, informality, diversity, and complexity. In these terms, Table 2 summarises all the features necessary to analyse the message-based factor of news text. After the feature extraction phase, we make use of the XGBoost (Extreme Gradient Boosting) algorithm [35] to compute the message-based value $M_i \in [-1, 1]$. Concerning other classifier algorithms, XGBoost does not present issues with poorly cured datasets and allows us to find out which features are more dominant and important for classification.

4.2 Fact-based

The Fact-based factor ensures the correctness of the news information and analyses its level of truth. It concerns two different contributions: the parameter $F_{i,f}$, which evaluates the news by matching its claims with a pre-trained network, named FEVER, of verified claims, and the Fact Comparison parameter $F_{i,c}$, which compares the news article with similar ones and uses this comparison for its evaluation.

4.2.1 FEVER

This parameter is responsible for the examination of facts by matching the text news with a pre-trained network based on Wikipedia claims, namely FEVER [36]. The network is able to classify a specific claim into three categories: supports, refutes, and not enough information. The first category refers to approved claims, i.e., the algorithm has found a correlation with the Wikipedia dataset. The second one considers false claims, which means there is evidence in the dataset that prove the claim is false, while The final category indicates that the network is unable to find any evidence regarding the reliability of the claim in question.

At first, the news text is processed and split into its different claims; we then define a set of claims $\mathcal{C}_i = \{c_{p,i}\}$ associated with news n_i. The generic claim $c_{p,i}$ can have two states: 1 for true claims that align with the FEVER dataset, and 0 for false or insufficient information claims. Therefore, the number of supported claims V_i is expressed as follows:

$$V_i = \sum_{p=1}^{|\mathcal{C}_i|} c_{p,i} \tag{2}$$

However, the number of reliable claims is highly dependent on the length L_i of the news, so naturally, longer news has more verified claims w.r.t. shorter news even if the news itself is not necessarily true. To this, we normalise the number of verified claims with the news length, i.e., V_i/L_i, and use this ratio, which values are in the interval $[0, 1]$, as an indicator of the reliability of the news. As shown in Figure 3, the number of verified claims, obtained from the Kaggle dataset, which will be explained in detail in Section 6, grows much slower when compared to the length of the news, so we expect that fake news will have a higher value of V_i/L_i w.r.t. real news. The FEVER parameter $F_{i,f}$ is then computed as:

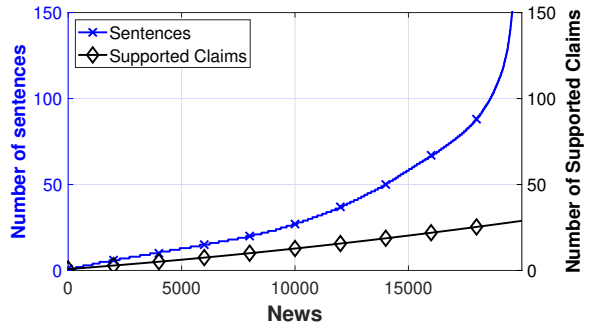


Fig. 3: Analysis of the number of supported claims and news length for the Kaggle dataset.

$$F_{i,f} = \begin{cases} 0.5 - \frac{V_i}{L_i} & \text{for } V_i > 0 \\ -0.5 & \text{for } V_i = 0 \end{cases} \quad (3)$$

so that $F_{i,f} \in [-0.5, 0.5]$ to account for the incompleteness of the FEVER dataset.

4.2.2 Fact Comparison

This parameter evaluates the accuracy of the facts reported in the news by comparing them to reliable and previously evaluated news, which are selected using a similarity algorithm. Since ground-truth information is unavailable for every news item, this approach provides a means for assessing news precision. We make use of the similarity approach presented in [37] in order to find comparable news articles that were already evaluated and stored in the database. At first, the considered news is processed, and a bag of words is generated through the combination of several embedding algorithms, such as Word2vec or GloVe. Then, the model has access to the database and retrieves similar already evaluated news with a higher value of feedback. Therefore, the news n_i is associated with a set of similar news $\mathcal{D}_i = \{n_j : f_j \geq f_{th} \;\&\; A_{i,j} > A_{min}\}$, where f_{th} represents the threshold beyond which the news is classified not only as information but also as reliable, $A_{i,j} \in [0, 1]$ depicts the similarity coefficient between two news n_i and n_j, and finally A_{min} is the minimum acceptable value of similarity. In specific, values of $A_{i,j}$ close to 1 refer to highly similar news, while 0 refers to completely different ones. For each similar news, the model associates its specific set of words, and so the similarity algorithm, namely the cosine similarity, is adopted to compare them and find the closest news. The fact comparison parameter $F_{i,c} \in [-1, 1]$ is then calculated only if there is similar news, i.e., if $|\mathcal{D}_i| > 0$, as follows:

$$F_{i,c} = \frac{\sum_{j=1}^{|\mathcal{D}_i|} f_j \cdot A_{i,j}}{\sum_{j=1}^{|\mathcal{D}_i|} A_{i,j}} \quad (4)$$

where f_j represents the feedback of similar news n_j.

Finally, the fact-based factor $F_i \in [-1, 1]$ for a news n_i is computed as follows:

$$F_i = \begin{cases} F_{i,c} & \text{if } |\mathcal{D}_i| > 0 \\ F_{i,f} & \text{if } |\mathcal{D}_i| = 0 \end{cases} \quad (5)$$

where if the number of similar news is equal to 0, so $|\mathcal{D}_i| = 0$, the factor is calculated based on the FEVER parameter; otherwise, the fact comparison parameter is used to evaluate the credibility of the news claims.

4.3 Context-based

The context-based factor takes care of all the parameters that are not directly related to the news but involve its context. In specific, it concerns the presence of ads on the web page $C_{i,a}$ and the sentiment analysis $C_{i,s}$. These two parameters are described below:

4.3.1 Advertisements

Low-credibility news sites usually make use of ads to gain significant revenue by attracting users [38]. For this reason, the system proposes to detect the common pattern of ads and take advantage of the correlation between fake websites and the number of ads as follows:

$$C_{i,a} = \begin{cases} 1 - 2\frac{N_{i,a}}{N'_{i,a}} & \text{for } N_{i,a} \leq N'_{i,a} \\ -1 & \text{for } N_{i,a} > N'_{i,a} \end{cases} \quad (6)$$

with $C_{i,a} \in [-1, 1]$. $N_{i,a}$ depicts the number of ads detected on the news web page, while $N'_{i,a}$ represents the maximum value after which every number of ads corresponds to the lowest value of the score, that is -1.

4.3.2 Sentiment Analysis

This parameter has the important role of detecting discrepancies between the sentiment related to the news text and its title, and it analyses the global text objectivity. The sentiment analysis determines if the information of the two components, text and title, is expressed in a positive, neutral, or negative way [39], and, in addition, it depicts the sentiment polarity, i.e., the strength of negative or positive sentiments [40]. Usually, fake content mixes different information with positive or negative feelings to mislead readers. Moreover, subjective language is commonly exploited by fake providers that focus on personal interpretation rather than factual data from an objective point of view [41]. The proposed model makes use of the sentiment algorithm suggested in [42] in order to analyse the sentiment related to the news title $S_{i,title}$ and the news text $S_{i,text}$; moreover, it takes advantage of a well-known sentiment analyser illustrated in [43] to evaluate the text objectivity. Therefore, our sentiment parameter presents two contributions: the first refers to the dissimilarities between text and title sentiment, which indicate if the title is coherent with the reported news and is not only a clickbait; meanwhile, the second one evaluates the global objectivity of the news text. The overall sentiment analysis factor $C_{i,s} \in [-1, 1]$ is then calculated as follows:

$$C_{i,s} = \sigma(2O_i - 1) - (1 - \sigma)(2\Delta S_i - 1) \quad (7)$$

where $\sigma \in [0, 1]$ depicts the weight selected to give more importance to a specific parameter and $O_i \in [0, 1]$ represents the level of objectivity; the value of 0 refers to a very subjective text, while 1 to a completely objective point of view. Moreover, the sentiment distance parameter $\Delta S_i = |S_{i,title} - S_{i,text}|$ measures the difference between

the sentiments expressed in the news title and text, where $\Delta S_i \in [0][1]$ and $S_{i,title}, S_{i,text} \in [0][1]$.

Finally, the context-based factor $C_i \in [-1, 1]$ for a news article n_i is expressed as follows:

$$C_i = \rho C_{i,a} + (1 - \rho) C_{i,s} \tag{8}$$

where $\rho \in [0, 1]$ is selected to give more importance to a specific contribution. Values of C_i close to -1 indicate fake content, while positive scores near 1 refer to real ones.

5 Trust Model

According to the scenario presented in Section 3, the trust model is designed to evaluate the news producers and estimate their credibility. Therefore, the model accesses the database to retrieve all relevant information about a specific news source, including news feedback and topic. When the system evaluates a source s on a specific topic t, it calculates the trust value using the following formula:

$$T_s^t = \gamma E_s^t + \delta H_s^t + (1 - \gamma - \delta) G_s^t \tag{9}$$

All these addends are in the range $[0, 1]$ and the weights are selected to give more importance to a specific factor so that their global sum is equal to 1, in order to normalize the trust value in the interval $[0, 1]$. The trust value T_s^t is evaluated based on three novel factors: the *Expertise* E_s^t, the *Coherence* H_s^t and the *Goodwill* G_s^t.

As expressed in Section 3, the feedback related to specific news has multidimensional views and can be described using $f_{s,i}^t$, which is related to several parameters, including the news source s, the topic t, and the news item i. In Section 4, we used a simplified definition of f_i to provide a general description and illustrate its composition in detail. Below, we continue using the analysed notation and express the feedback as f_i for better reading. As a consequence, the definition of topic t and source s is omitted in all the parameters that are related to them, e.g., the trust factors described as follows: the *Expertise* E, the *Coherence* H and the *Goodwill* G.

5.1 Expertise

The first factor quantifies how the source is well-informed on a specific topic. In specific, the expertise factor E is evaluated based on two parameters: the Topic Importance E_m, which analyses the impact of the topic t in all the topics tackled by the provider s, and the Writing Competence E_c, which considers proficiency in writing news of the source on the evaluated topic.

5.1.1 Topic Importance

The first parameter measures the expertise of the source based on the number of news published on the evaluated topic and discriminates specialised and qualified providers from general ones. The Topic Importance $E_m \in [0, 1]$ is then computed as follows:

$$E_m = \frac{N_s^t}{\bigcup_t N_s^t} \tag{10}$$

where N represents the number of news belonging to the source s on the topic t, while $(\bigcup_t N_s^t)$ depicts the total number of published news by that source.

5.1.2 Writing Competence

The second parameter refers to the writing style computed as the average of the message-based evaluations, calculated according to Section 4.1. The Writing Competence E_c is then computed as:

$$E_c = \frac{1}{2N} \left(1 + \sum_{i=1}^{N} M_i \right) \tag{11}$$

where $M_i \in [-1, 1]$ represents the value of message-based factor of the news n_i published by the considered source s on the topic t.

Finally, the global Expertise $E \in [0, 1]$ is expressed as:

$$E = \tau E_m + (1 - \tau) E_c \tag{12}$$

where values of E close to 0 refer to sources with a low level of expertise, while scores close to 1 indicate very capable and expert providers. Moreover, the weight τ is used to give more importance to a specific contribution. Specifically, the weight gives more influence to the topic importance parameter with high numbers of published news, while the expertise on a topic is not relevant with only a few pieces of published news. In these terms, $\tau \in [0, 1]$ is expressed as follows:

$$\tau = \frac{N}{\frac{1}{\omega} N + \psi} \tag{13}$$

where the weights $\omega \in (0, 1]$ and $\psi \in [1, \inf)$ are used to set the asymptotic value of the weight and to configure its speed at the variation of N; a more detailed explanation of these weights will be provided in Section 6.3.

5.2 Relevance and Goodwill

These last two factors are used to study the dissemination of the source's news among the users and to take advantage of the social impact of fake news in the detection mechanism. Several research models demonstrate the risk of fake content spreading in social networks and how the perceived information quality is influenced by the intention to re-share information [44]. In these terms, we propose the Relevance factor $R_i \in (0, 1]$ for a generic news article n_i to determine how frequently a news source's articles are consulted and to weigh its evaluation accordingly; the factor is expressed as:

$$R_i = \frac{N_{r,i}}{N_r} \tag{14}$$

where $N_{r,i}$ represents the number of times the news n_i is requested by users, N_r defines the total number of news requests for the source s within the topic t, and $\left(\sum_{i=1}^{N} N_{r,i} = N_r\right)$. The news feedback f_i is then weighted through the Relevance factor and the global Goodwill factor $G \in [0, 1]$ is computed as follows:

$$G = \frac{1}{2} \left(1 + \sum_{i=1}^{N} R_i \cdot f_i \right) \tag{15}$$

where R_i and f_i represent the Relevance factor and the feedback related to the news n_i, respectively.

TABLE 3: Simulation Parameters

Parameter	Description	Value
f_{th}	Minimum feedback threshold to consider for the "fact-comparison"	0.8
A_{min}	Similarity threshold	0.86
$N'_{i,a}$	Ads threshold	2
σ	Sentiment Analysis parameter	0.6
ω	Expertise asymptotic parameter	0.3
ψ	Expertise speed parameter	70
p	Parameter of the geometric distribution for the Coherence	0.03
N_{last}	Temporal Window for assessing the source's trust score	50
γ	Expertise weight	0.25
δ	Coherence weight	0.5

5.3 Coherence

The Goodwill factor may not be effective in responding to sudden changes in a source's behaviour, as it happens for dynamic attacks such as On-Off Attacks (OOA) [45] and Whitewashing Attack (WA) [46]. Indeed, sources that implement these attacks periodically change their behaviour, e.g., by alternatively being benevolent (ON) and malevolent (OFF) or by registering again as news providers with a different identity. To overcome these attacks, this factor evaluates the recent behaviour of a source considering a small temporal window, which makes use of the last N_{last} news evaluated for source s on topic t. The Coherence factor $H \in [0, 1]$ is then computed as follows:

$$H = \frac{1}{2}\left(1 + \sum_{z=1}^{N_{last}} w_z \cdot f_z\right) \tag{16}$$

where f_i depicts the feedback of the news n_i and N_{last} represents the dimension of the temporal windows of considered feedbacks. To give more relevance to the latest feedback, with regard to the oldest one, the weight w_z of each feedback follows a geometric distribution with parameter p:

$$w_z = p(1-p)^{z-1} + \frac{\xi^{res}}{N_{last}} \tag{17}$$

where to maintain the parameter range $[0, 1]$, we introduce the term ξ^{res}, which accounts for all the residual weights of the distribution due to the transactions older than N_{last}. Therefore, ξ^{res} is computed as follows:

$$\xi^{res} = \sum_{r=N_{last}+1}^{N} p(1-p)^{r-1} \tag{18}$$

6 Experimental Evaluation

6.1 Simulation Setup

In order to test the proposed system, i.e., both the fake news detector and the trust management algorithm, we need a large dataset of news. To this, we make use of the dataset made available by Kaggle at [47], which contains a total of 20387 news. Kaggle is a platform that hosts data science competitions and organizes tournaments for recruitment and academic research. In detail, the dataset is in a Comma-Separated Values (CSV) format and consists of four attributes: title, author, text, and label. The title is the headline of the published news; the author represents the journalist who wrote the news; the text is the body content that describes in detail the news story and, finally, the label attribute defines, through a binary classification of zero or one, whether the news is false or real, respectively.

Given the large size of the dataset, we consider that out of all the articles, 90% of them are used for training the message-based factor, whereas the remaining 10% was used to test its performance. However, a preprocessing phase was required to manage the data for our purposes properly; first, we adjusted the ground truth labels so that fake news is labelled with -1, while genuine news articles have a label 1. Second, the system required the news source, which is used by the trust evaluation model: thus, the author field was discarded and replaced by the news article URL. Another effort for retrieving the necessary attributes to the overall system, which was not comprised of the original dataset, was the main topic covered by the news. To retrieve this information, we implemented a deep learning approach to detect it automatically. The purpose of this network is to classify each news article into one of 10 possible topics: Arts & Culture, Business & Economy, Crime & Security, Entertainment & Celebrity, Health & Education, Politics, Science, Sports, Tech, and finally, Weird. This classification was obtained using a dataset of 60,000 news items gathered from various news producers' websites with web scraping techniques. In particular, the dataset consists of two attributes: news' text and category. The news text is cleaned and embedded in order to train the network. Therefore, a percentage of 80-20% was used to divide the dataset among training and testing using the news embeddings as the feature vector. The results were satisfactory and precise for our purpose: we achieved a test accuracy of 91%.

Finally, the dataset only classifies news as fake or real, so to test our detector algorithm, we consider that news with feedback lower than 0 is fake, and real otherwise. Table 3 shows the optimal configuration of the simulation parameters for the proposed system and the different weights used for the model. More details on the selection of these parameters are given in the following Sections.

6.2 Simulation Results for the Feedback Model

The feedback model aims to evaluate news from various perspectives and aggregate them in order to provide a feedback score that represents the reliability of a specific news article. Therefore, the feedback score f_i related to news n_i is a proper combination of metrics based on the message, fact, and context parameters, in which values near -1 mean n_i is likely fake news, i.e., it has the temperament of disinformation. On the other hand, values around 1 suggest that the feedback model detects news that provides information.

6.2.1 Message-based

The first set of simulations aims to validate the performance of the XGBoost algorithm and to compare it with other four machine learning algorithms, Naive Bayes (NB) [48], Support Vector Machine (SVM) [49], K-nearest neighbours (KNN) [50] and Random Forest (RF) [51]. In order to evaluate the classifiers' performance, we used two well-known

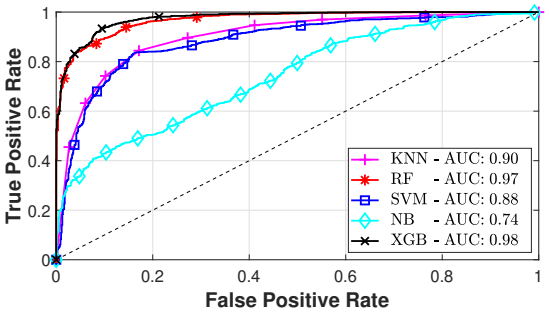


Fig. 4: ROC curves for the proposed machine learning classifiers.

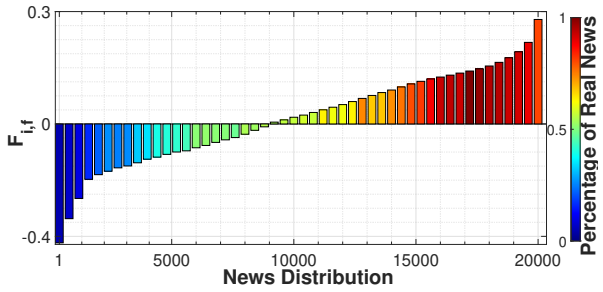


Fig. 5: FEVER Score distribution

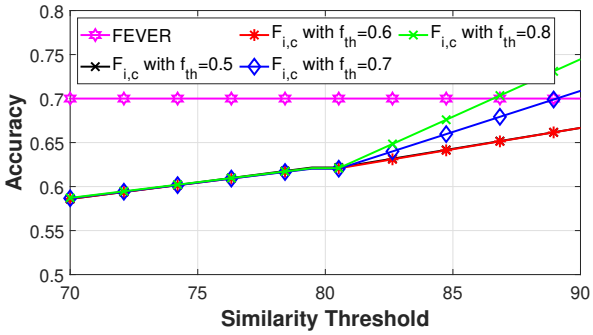


Fig. 6: Fact-based overall accuracy over different thresholds of feedback and similarity

metrics: the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). The ROC curve measures the true and false positive rates at different classification thresholds. Alternately, the AUC illustrates how the model is accurate in achieving the classification through the output classes, e.g., AUC=1 means that the system is 100% accurate. A comparison of the proposed classifiers is shown in Figure 4, where the XGBoost and RF emerge as the most appropriate choice. Among these two best-performing algorithms, we decided to use XGBoost, which presents the best results in terms of computational speed in our simulations, with a total training time of less than a minute. With this setup, the message-based factor reached a 92% accuracy, with 91% concerning fake and 93% for real news.

6.2.2 Fact-based

The focus of the following set of simulations concerns the testing of the fact-based factor, i.e., the combination of the contributions from FEVER, which has been tested to retrieve the number of supported claims, and from the fact comparison with already evaluated news.

We decide to consider the number of supported claims for news due to the study of [52], which reveals how FEVER improves its accuracy in detecting only supported claims. At this point, the Stanford Parser tool [53] was used to decompose each news article into triplets, i.e., subject, verb, and object, and to remove unnecessary parts of the speech, so as to provide a simplified input for FEVER. Once the data was fitted properly, FEVER was employed to compute the number of supported claims V_i for news n_i. Figure 5 shows the FEVER contribution $F_{i,f}$ by displaying the percentage of real news within each bar with different colour shades. For a clearer view, each bar reflects an aggregation of 400 news. As expected, high values of V_i/L_i indicate a high concentration of fake news, whereas low values imply a prevalence of real news. Additionally, the more the FEVER parameter is close to the boundaries, the more its accuracy increases. Finally, the FEVER parameter reaches an overall accuracy of 70%, computed by assessing fake news if $F_{i,f} < 0$ and real news if $F_{i,f} \geq 0$.

The second contribution of the fact-based module is the fact comparison, which employs the concept of similarity. As explained in section 4.2.2, the concept of similarity is exploited in order to achieve a comparison with previously analysed news. In detail, we operate on two parameters: the similarity threshold A_{min} and the minimum feedback threshold f_{th}, to consider only a small group of similar and reliable news already evaluated.

Figure 6 shows the fact-comparison accuracy scores (compared with the steady trend of FEVER) by varying the similarity and the feedback thresholds. The highest accuracy value can be obtained using stringent thresholds, i.e., by selecting news with a 90% of similarity and feedback of at least 0.8; however, these thresholds are too demanding, and only the 5% of news could be evaluated with them. By loosening the thresholds slightly, i.e., by setting $A_{min} = 0.87$, we were able to include the 23% of news and obtain an accuracy better than FEVER. For all the other news, which can not be evaluated with the fact comparison, the system returns the value computed by the FEVER contribution.

6.2.3 Context-based

The set of simulations on the context is divided into two categories: the analysis of the ads available on the news web page and the sentiment analysis. Figure 7 depicts the impact of advertisements on the news classification; indeed, real news exhibit only one or two ads and it is clear how after a certain number of ads, denoted by $N'_{i,a}$ in our model, greater than 2, the presence of real news drastically decreases. This allows the system to provide a completely negative feedback score, i.e., $C_{i,a} = -1$, if $N_{i,a} > 2$. Furthermore, we analyse the importance of sentiment analysis. Table 4 shows the mean and variance of the sentiment for the title and news text as well as their difference ΔS, and the mean and the variance of the text objectivity. The results demonstrated that fake news tends to have more divergent values, with a greater average and variance. The system takes advantage

TABLE 4: Investigation for the sentiment analysis.

	Real		Fake	
	$\overline{x}$	σ^2	$\overline{x}$	σ^2
S_{title}	0.78	0.011	0.71	0.022
S_{text}	0.65	0.075	0.55	0.085
ΔS	0.21	0.036	0.25	0.041
O	0.41	0.007	0.42	0.017
$C_{i,s} = 0.72$				

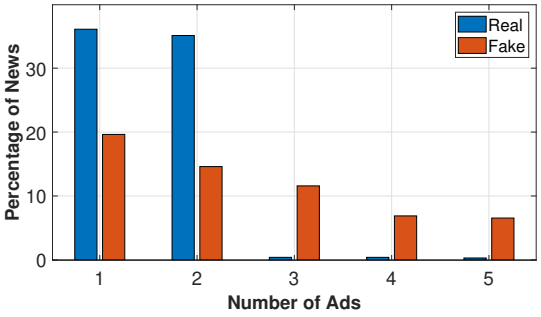


Fig. 7: Impact of advertisements (ads)

of these differences and increases the accuracy of fake news detection so that the final accuracy of the context-based factor is 76%.

6.2.4 Feedback Score Overall

This Section concerns the final aggregation of the three factors analysed above: message-based, fact-based and context-based. The message-based contribution is the most accurate, so the weights are chosen to provide more importance to this factor. The best results are obtained by setting $\alpha = 0.7$ and $\beta = 0.15$, which resulted in the same accuracy as the message-based factor alone. Indeed, giving more weight to the fact-based or context-based factor resulted in lower accuracy. However, we decided to investigate each factor's ability to detect fake news, so we analysed the accuracy of each factor at different intervals employing an additional dataset to generalise the weights' choice. The new dataset has been collected by Ahmed et al. [54] with almost 40,000 news articles. The results are reported in Table 5. From this analysis, we can notice that the message-based factor is less accurate in those intervals, which are close to zero. In these intervals, fake news behaves similarly to factual news in terms of detailed information and writing quality. From these considerations, we use the term "misinformation" to describe these areas of significant ambiguity. The new term does not necessarily fit into the solution's metrics, although the term "misinformation" follows the literature and explains this area of uncertainty. However, in terms of accuracy, we distinguish only fake and real news by checking whether the feedback is respectively lower or greater than zero. Hence, due to the uncertainty intervals of the message-based factor, the weights can be adjusted to exploit its low accuracy.

This means that α and β have no constant values but rather their value changes to take advantage of the strong point of each factor to improve the overall accuracy. In order to allow only the factors with good accuracy to provide a contribution in the final aggregation, the weight for all

TABLE 5: System's factors accuracy for the two analyzed datasets.

Interval	Message-based Accuracy	Fact-based Accuracy	Context-based Accuracy
Kaggle Dataset [47]			
[-1,-0.5]	0.95	0.74	0.86
(-0.5,0]	0.62	0.7	0.84
(0,0.5]	0.7	0.76	0.3
(0.5,1]	0.96	0.7	0.7
Ahmed et al. Dataset [54]			
[-1,-0.5]	0.96	0.72	0.73
(-0.5,0]	0.64	0.69	0.69
(0,0.5]	0.61	0.77	0.75
(0.5,1]	0.96	0.73	0.78

TABLE 6: Accuracy comparison among different related works with the same dataset.

Work	Accuracy
Drif et al.	0.725
Ahmed et al. [LR-Unigram]	0.89
Ahmed et al. [LR-LSVM]	0.92
FAKE	0.94

factors with accuracy less than 0.5 is set to 0. We, therefore, need to re-scale the accuracy from $[0.5, 1]$ to $[0, 1]$. To this, we are interested in assigning greater weight to the factors that have higher accuracy values, so we have adopted a non-linear transformation and, in particular, a square function with the vertex in $(0.5, 0)$ and passing through the point $(1, 1)$. Finally, we normalise the obtained accuracies so that their sum is equal to 1, to obtain the weights for the three factors. We then tested these weights for the two databases proposed, and we achieved a 94% accuracy for both the Kaggle dataset and the dataset proposed by Ahmed et al. Finally, we remark that even though the overall accuracy achieved is only 2% higher than message-based accuracy, the fact and context-based accuracy provide a 25% increase in the accuracy in terms of remaining errors.

In conclusion, a comparison with previously studied authors who employed the same Kaggle dataset as our solution is provided. Specifically, we tested our algorithm on the above-mentioned dataset by keeping the exact same number of news used by related works in order to have a correct comparison among them. Drif et al. [17], through multiple approaches such as content-based, user-based and social-based, achieved the lowest accuracy of 72.5%. Another significant work is covered by Ahmed et al. [16], in which they tested two methods: the first one with an LR-unigram and the second one with LR-LSVM obtaining an accuracy of 89% and 92% respectively. Finally, another similar approach through the same dataset was employed by [19], using a Naïve Bayes classifier. They split the dataset into 70% for training and 30% for the test part and obtained 92% accuracy, which is perfectly comparable with our message-based performance. Table 6 summarises the performance of these works and compares them with our solution.

6.3 Simulation Results for the Trust Model

We evaluate the performance of the proposed trust model by analysing the trust value. A fine-tuning of the weights related to the three metrics, i.e., Expertise, Goodwill, and Coherence was necessary. The most relevant results have

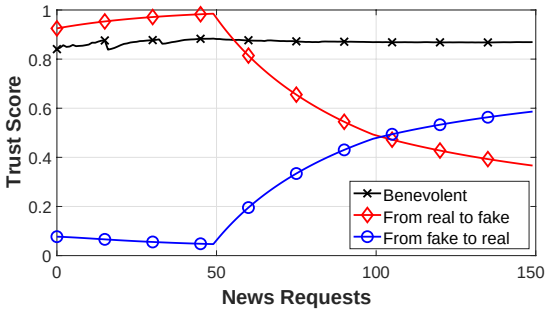


Fig. 8: Final trust score: Three different behaviours.

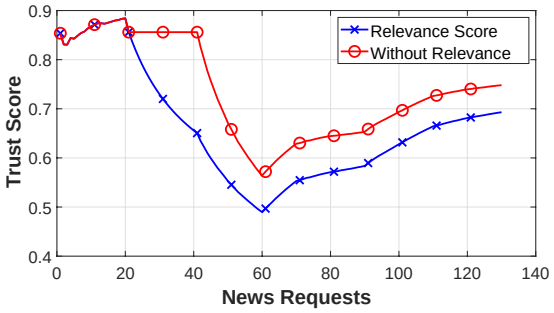


Fig. 9: Trust score with and without the relevance algorithm.

been achieved by setting $\gamma = 0.25$ and $\delta = 0.25$. Figure 8 depicts the trust value with these last parameters when a source performs three distinct behaviours. In general, media outlets could modify their behaviours for several reasons; this happens for the competitive nature of journalism, which can lead to reporters feeling pressured and publishing news as soon as possible without verifying its authenticity [55]. To test it, we adopt three different behaviours: a benevolent one, in which the news provides verified and real information, and two dynamic ones, aimed at testing the robustness of our system to news media that modify their behaviours after publishing either real or fake news. In the first dynamic behaviour, the source builds its reputation with 50 trustworthy news and then starts providing 100 fake news, while in the second one, the source begins publishing fake news and then tries to improve its reputation with 100 reliable news after having provided 50 fake news. The simulations show how the algorithm is able to quickly adapt to the changes thanks to the Coherence factor.

The following results focus on a better understanding of how the trust model avoids the spreading of fake news. To this, we have organized simulations with synthetic data: firstly, an experiment was conducted to build an attack for relevant news. The first group of 20 real news is induced to bring the algorithm to convergence. Then, a single piece of fake news is requested 20 times, followed by 20 different fake news requests; finally, 70 real news are requested to see how the system reacts.

Figure 9 depicts the system's response to these attacks in two cases: with and without the relevance factor. It is clear that once the source publishes the first 20 real news, the trend is the same for both algorithms. The discrepancies occur when the same fake news is required multiple times,

TABLE 7: Trust and average feedback scores related to the most evident sources

Source	Topic	% Fake news (number of fake, number of real)	Average Feedback, Trust score
NYTimes	B&E	1.25 (1, 80)	0.94, 0.92
Breibart	B&E	0 (0, 86)	0.76, 0.73
NYTimes	Politics	0 (0, 170)	0.94, 0.91
Breibart	Politics	0,54 (1, 185)	0.72, 0.69

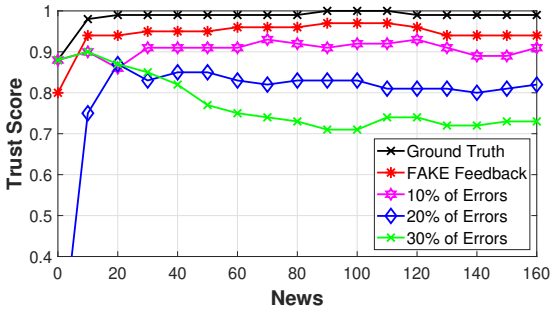


Fig. 10: Trust model performance through different fake news detector accuracy systems.

making the system comprehend that the news article is significant. The trust score rapidly decreases when the algorithm detects news requested several times, and so it penalizes the source. In contrast, when relevance is not employed, the news provider is punished only for the single fake news, and the trust score is steady until the source publishes 20 more different fake news. Moreover, at this point, we can notice that also the trust value of the source with relevance changes behaviour due to the different penalization: indeed, at first, the negative contribution was due to the multiple requests of the same fake news already stored in the database. However, the curve decreases rapidly due to the bigger impact of new fake news evaluation. In addition, during the burst of 70 real news, both curves are able to regain a part of their trust score. Another interesting result can be noticed during the rising edge: two points present abrupt changes due to the Coherence factor, employing a short temporal window N_{last} to assess the recent news. When the temporal windows N_{last} contain positive feedback scores, the trust value increases more rapidly.

Another significant achieved result is described in Table 7. The simulation is based on the real dataset that includes sources that primarily write on the topics of Business & Economy (B&E) and Politics. The table shows the quantity of fake and true news stories for each source and topic. Furthermore, the average feedback is reported, followed by the source's trust score on that specific subject. A considerable result is given from the fact that, although the source "Breitbart" has a lower percentage of news labelled as fake by our fake news detector than the source "NYTimes", the latter receives a higher score. This is because our system not only detects fake news but assesses its quality. Indeed, a news story with a feedback score close to 1 is well-written due to the design of the feedback evaluation score, which contains continuous values and is provided by the high impact of the message-based.

Finally, we want to understand how the accuracy of

TABLE 8: Machine Specifications

	Client (Plugin)	Server (FAKE System)
Processor	Intel(R) Core(TM) i7-5500U CPU @2.40GHz	Intel(R) Xeon(R) E5-1620 CPU @3.60GHz
Memory	8GB Samsung 1600MHz	32GB Samsung 1600 MHz
Storage	Samsung SSD 870 QVO 1TB	SanDisk SSD PLUS 240GB
Operating System	Windows 10 Home 64 bit	Ubuntu 22.04 LTS

TABLE 9: Average impact of the plugin over 100 users.

	No-Plugin	Plugin
Average Time Execution [s]	1.23	1.54
Average Memory Usage [MB]	25.74	79.12

detecting fake news affects the trust model evaluations. Figure 10 represents a reliable source that publishes news. The ground truth case (black line) has been highlighted, in which the fake news detector is 100% accurate. In this case, the feedback evaluation concerns a discrete rate where fake news has a feedback score of $f_i = -1$, while $f_i = 1$ is assigned to real news. In contrast, the red line represents the trust score obtained by the feedback computed by the proposed fake detector. In addition, we want to analyze the results at varying the error percentage in the feedback evaluation. The other curves show how the trust model reacts with a 10, 20, and 30% error, respectively, and how it can follow the real trend. Although the feedback model is not 100% accurate, the trust model manages the errors quite well by following the ideal trend until an error of 20%.

6.4 Simulation Results for the Entire System

This last section concerns the functioning of the entire system in a real environment. The system is tested by simulating the activity of multiple users looking for news on a web browser for websites belonging to three well-known media sources, i.e., BBC, Breitbart, and USA Today. In these terms, Table 8 illustrates the specific of the machines used for the following simulations, where the Client indicates the machine used by users and in which the plugin is installed. The Server is running the FAKE system, responsible for all the evaluations. We evaluate the performance of the system by testing 100 simultaneous users, in which all the plugins (clients) detect news and send them to the server for evaluation. The results are analysed in terms of the time processing necessary for the server in order to perform evaluations. In this regard, the system is able to evaluate news in less than an average of 2 seconds and is able to provide the evaluations for news already evaluated in less than 0.2 seconds. Furthermore, we analyse the impact of the plugin on browser performance by measuring the Average Time Execution, which represents the average time needed for displaying the evaluations in browsers, and the Average Memory Usage, which is the average amount of memory, usually expressed in bytes, required to load the data. Table 9 illustrates the average impact of the plugin by analysing the browser without and with the usage of the plugin. Simulations illustrate the low effects of the plugin in terms of time and memory usage.

TABLE 10: Database and Blockchain comparison.

Database	**Write**			**Read**		
# News	1	500	1000	1	500	1000
Latency [s]	0.0019	0.0036	0.0054	0.00046	0.00042	0.00048
THR [MBps]	31.45	24.20	18.03	647.90	714.15	693.53
Blockchain	**Write**			**Read**		
# News	1	500	1000	1	500	1000
Latency [s]	1.20	10.04	20.44	0.080	0.077	0.078
THR [MBps]	0.30	0.06	0.032	4.46	4.37	4.40

7 Secure Data Management: Blockchain Technology

This section compares *Blockchain* technology [56] with traditional databases. Blockchain can be implemented as an alternative to a secure database, and it has been used to save all feedback and source evaluations [57], [58]. In recent years, the Blockchain has gained massive popularity in many research areas, with several approaches proposed for using it to combat fake news and provide a transparent and secure environment.

Blockchain can be described as a digital ledger of transactions, duplicated and distributed across a decentralised ecosystem, enabling trust in peer-to-peer networks without the presence of certification authority. In contrast to centralised systems, Blockchain has overcome several security weaknesses, such as being tampered with by malicious actors. Furthermore, due to the distributed ledger maintained using the distributed consensus algorithm, the Blockchain enables the involved actors to avoid third-party trust in interactions, guaranteeing traceability and security [59]. More abstractly, Blockchain can be seen as an ordered list of blocks, where each block represents a register that keeps information and transactions, and it is linked to the previous one in chronological order. The chain starts with the genesis block, representing the first block. All the information in all blocks is encrypted, and a consensus algorithm ensures reliability for new nodes in the network. For this comparison, we have implemented our system, making use of both an encrypted database, as described in the rest of the paper, and the Ethereum Blockchain proposed by Alastria [60]. In the latter case, each news evaluation provided by the detector is stored in the Blockchain, and a new block is created; this process guarantees that any modifications can not be achieved without changing all the previous blocks. Moreover, the Blockchain provides real-time responses to the request of news that are already evaluated and are necessary for the credibility measure of the news providers.

In an Ethereum Blockchain, the smart contracts allow the users to execute a script on a Blockchain network in a verifiable, way and so create and retrieve transactions using private functions. An Ethereum node can follow the instructions in a smart contract and execute them from a valid account. The contract determines the structure of the data that can be stored in the Blockchain and the functions with which the nodes operate within the Blockchain. In this solution, we wrote a smart contract making use of the Solidity programming language. Furthermore, in order to compare the Blockchain with the encrypted database, Table 10 shows a set of simulations for both actions, i.e., write and read. We evaluate the performance by analysing two

metrics: latency, which specifies the time taken to read or write data, and throughput, which represents the amount of information that can be written or read in a given amount of time, typically measured in bits per second (bps). The Table illustrates that the encrypted database outperforms the Blockchain in both metrics, and it exhibits better performance in terms of processing time, making it the preferred solution for real-time applications. Moreover, one ongoing challenge in Ethereum is scalability, given that the platform currently processes approximately 500,000 transactions per day and has a maximum capacity of about 15 transactions per second [61]. However, in scenarios where trust, robustness, and data provenance are the system's top priorities, Blockchain is the best solution and overcomes several security weaknesses, such as tampering by attackers [62].

8 Conclusions

In this paper, we have proposed a detection algorithm for text-written news, classifying articles as real or fake according to several parameters related to the text style and the news context. Moreover, we have developed a trust and reputation system so that it is possible for users to understand which news providers that can lead to successful collaboration, i.e., that can provide reliable news. The paper also presents a section to compare traditional databases with a Blockchain solution based on the Ethereum smart contract. Finally, we tested the whole system using a Kaggle dataset containing several articles from various domains.

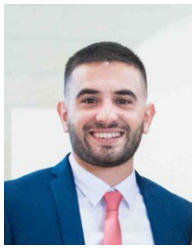
Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 957228.

References

[1] B. Kalsnes, "Fake news," in *Oxford Research Encyclopedia of Communication*, 2018.
[2] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.
[3] S. Vosoughi *et al.*, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
[4] S. Lewandowsky and S. Van Der Linden, "Countering misinformation and fake news through inoculation and prebunking," *European Review of Social Psychology*, vol. 32, pp. 348–384, 2021.
[5] W. T. EU, "Trust and reputation systems in blockchain technology," 2020.
[6] K. Sharma *et al.*, "Combating fake news: A survey on identification and mitigation techniques," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 3, pp. 1–42, 2019.
[7] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.
[8] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends in cognitive sciences*, vol. 25, no. 5, pp. 388–402, 2021.
[9] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, pp. 38–55, 2019.
[10] J. Kim *et al.*, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation," in *Proceedings of the eleventh ACM international conference on web search and data mining*, 2018, pp. 324–332.
[11] R. Mujumdar and S. Kumar, "Hawkeye: a robust reputation system for community-based counter-misinformation," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021, pp. 188–192.
[12] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.
[13] X. Li *et al.*, "An empirical study on how well do covid-19 information dashboards service user information needs," *IEEE Transactions on Services Computing*, 2021.
[14] V. Singh *et al.*, "Automated fake news detection using linguistic analysis and machine learning," in *International conference on social computing, behavioral-cultural modeling, & prediction and behavior representation in modeling and simulation*, 2017, pp. 1–3.
[15] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment aware fake news detection on online social networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2507–2511.
[16] H. Ahmed *et al.*, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*. Springer, 2017, pp. 127–138.
[17] A. Drif, Z. F. Hamida, and S. Giordano, "Fake news detection method based on text-features," in *The Ninth International Conference on Advances in Information Mining and Management*, 2019.
[18] A. Pathak and R. K. Srihari, "Breaking! presenting fake news corpus for automated fact checking," in *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*, 2019, pp. 357–362.
[19] F. I. Adiba *et al.*, "Effect of corpora on classification of fake news using naive bayes classifier," *International Journal of Automation, Artificial Intelligence and Machine Learning*, vol. 1, pp. 80–92, 2020.
[20] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," *Procedia Computer Science*, vol. 141, pp. 215–222, 2018.
[21] R. Vijjali *et al.*, "Two stage transformer model for covid-19 fake news detection and fact checking," in *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2020, pp. 1–10.
[22] N. Vo and K. Lee, "The rise of guardians: Fact-checking url recommendation to combat fake news," in *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018, pp. 275–284.
[23] B. Bhutani *et al.*, "Fake news detection using sentiment analysis," in *2019 twelfth international conference on contemporary computing (IC3)*. IEEE, 2019, pp. 1–5.
[24] T. Gröndahl and N. Asokan, "Text analysis in adversarial settings: Does deception leave a stylistic trace?" *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.
[25] Z. Epstein, G. Pennycook, and D. Rand, "Will the crowd game the algorithm? using layperson judgments to combat misinformation on social media by downranking distrusted sources," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–11.
[26] R. Wang *et al.*, "Rumorlens: Interactive analysis and validation of suspected rumors on social media," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–7.
[27] S. Tschiatschek *et al.*, "Fake news detection in social networks via crowd signals," in *Companion proceedings of the the web conference 2018*, 2018, pp. 517–524.
[28] C. Fisher, "What is meant by 'trust'in news media?" in *Trust in media and journalism*. Springer, 2018, pp. 19–38.
[29] F. Prochazka and W. Schweiger, "How to measure generalized trust in news media? an adaptation and test of scales," *Communication Methods and Measures*, vol. 13, no. 1, pp. 26–42, 2019.
[30] J. Strömbäck *et al.*, "News media trust and its impact on media use: Toward a framework for future research," *Annals of the International Communication Association*, vol. 44, pp. 139–156, 2020.
[31] J. D. Martin and F. Hassan, "News media credibility ratings and perceptions of online fake news exposure in five countries," *Journalism Studies*, vol. 21, no. 16, pp. 2215–2233, 2020.
[32] B. A. Buhari, A. A. Obiniyi, K. Sunday, and S. Shehu, "Performance evaluation of symmetric data encryption algorithms: Aes and blowfish," *Saudi Journal of Engineering and Technology*, vol. 4, pp. 407–414, 2019.
[33] K. Shu *et al.*, "Mining disinformation and fake news: Concepts, methods, and recent advancements," in *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 2020, pp. 1–19.
[34] D. Baines *et al.*, "Defining misinformation, disinformation and malinformation: An urgent need for clarity during the covid-19 infodemic," *Discussion Papers*, vol. 20, no. 06, pp. 20–06, 2020.

[35] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
[36] J. Thorne *et al.*, "The fact extraction and verification (fever) shared task," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 1–9.
[37] S. Agarwala *et al.*, "Detecting semantic similarity of documents using natural language processing," *Procedia Computer Science*, vol. 189, pp. 128–135, 2021.
[38] S. Castelo *et al.*, "A topic-agnostic approach for identifying fake news pages," in *Companion proceedings of the 2019 World Wide Web conference*, 2019, pp. 975–980.
[39] M. A. Alonso *et al.*, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, p. 1348, 2021.
[40] M. Thelwall *et al.*, "Sentiment strength detection in short informal text," *Journal of the American society for information science and technology*, vol. 61, no. 12, pp. 2544–2558, 2010.
[41] L. L. Vieira *et al.*, "Analysis of the subjectivity level in fake news fragments," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 2020, pp. 233–240.
[42] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.
[43] S. Loria *et al.*, "Textblob documentation," *Release 0.15*, vol. 2, p. 269, 2018.
[44] M. Koohikamali and A. Sidorova, "Information re-sharing on social network sites in the age of fake news." *Informing Science*, vol. 20, 2017.
[45] J. Caminha *et al.*, "A smart trust management method to detect on-off attacks in the internet of things," *Security and Communication Networks*, vol. 2018, 2018.
[46] R. Abassi, "Dealing with collusion attack in a trust-based manet," *Cybernetics and Systems*, vol. 49, no. 7-8, pp. 475–496, 2018.
[47] Kaggle, "Fake news," 2018. [Online]. Available: https://www.kaggle.com/c/fake-news
[48] H. Zhang and D. Li, "Naïve bayes text classifier," in *2007 IEEE International Conference on Granular Computing (GRC 2007)*. IEEE, 2007, pp. 708–708.
[49] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification*. Springer, 2016, pp. 207–235.
[50] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.
[51] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.
[52] Y. Nie *et al.*, "Combining fact extraction and verification with neural semantic matching networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6859–6866.
[53] M.-C. De Marneffe and C. D. Manning, "Stanford typed dependencies manual," Technical report, Stanford University, Tech. Rep., 2008.
[54] H. Ahmed *et al.*, "Detecting opinion spams and fake news using text classification," *Security and Privacy*, vol. 1, no. 1, p. e9, 2018.
[55] N. Fitzpatrick, "Media manipulation 2.0: the impact of social media on news, competition, and accuracy," 2018.
[56] D. Berdik *et al.*, "A survey on blockchain for information systems management and security," *Information Processing & Management*, vol. 58, no. 1, p. 102397, 2021.
[57] M. Saad *et al.*, "Fighting fake news propagation with blockchains," in *2019 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2019, pp. 1–4.
[58] Z. Shahbazi and Y.-C. Byun, "Fake media detection based on natural language processing and blockchain approaches," *IEEE Access*, vol. 9, pp. 128 442–128 453, 2021.
[59] L. Liu *et al.*, "A survey for the application of blockchain technology in the media," *Peer-to-Peer Networking and Applications*, vol. 14, no. 5, pp. 3143–3165, 2021.
[60] Alastria, "Alastria Blockchain Ecosystem," https://alastria.io/.
[61] R. K. Raman and L. R. Varshney, "Information-theoretic approaches to blockchain scalability," in *Handbook on Blockchain*. Springer, 2022, pp. 257–296.
[62] M. J. M. Chowdhury *et al.*, "Blockchain versus database: a critical analysis," in *IEEE International conf on trust, security and privacy in computing and communications*. IEEE, 2018, pp. 1348–1353.



Claudio Marche received the M.Sc. degree in telecommunication engineering with full marks in 2018 from the University of Cagliari. Since graduation, he has been working as Researcher in the Department of Electrical and Electronic Engineering at the University of Cagliari in the Net4U research group. He is currently a Ph.D. student in Electronic and Computer Engineering at the University of Cagliari. His current research interests include IoT, SIoT and Trust.

Ilaria Cabiddu received the M. Sc in electronic engineering with full marks in 2021 from the University of Cagliari. During the graduation, she has been working as Researcher in the Department of Electrical and Electronic Engineering at the University of Cagliari, in the MCLab research group. His current research interests include Internet of Things (IoT) and Artificial Intelligence.

Christian Giovanni Castangia : M.Sc graduated from the University of Cagliari in Internet Technology Engineering with full marks in 2022. Since 2020, he has been involved in research at the University of Cagliari, covering topics related to smart cities and smart mobility. Recently he has been working on topics related to machine learning and Blockchain.

Luigi Serreli is a Ph.D. student in Electronic and Computer engineering at the Department of Electrical and Electronics Engineering at the University of Cagliari (Italy). He holds a B.Sc. in Electrical and Electronic Engineering and an M.Sc. in Internet engineering at the University of Cagliari, discussing the thesis "A generative Adversarial Network (GAN) Fingerprint approach over LTE".

Michele Nitti is an Assistant Professor at the University of Cagliari, Italy since 2015. He served as a technical program chair for various international conferences (IEEE BMSB 2017, IEEE IoT V&T Summit 2020). Currently, he is a member of the editorial board for the IEEE IoT Journal, Elsevier Computer Networks and co-founder of an academic spin-off (GreenShare s.r.l.) which works in the mobility sector. His main research interests are in architecture and services for the Internet of Things (IoT).