IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media

NOUREDDINE SEDDARI[1,2], ABDELOUAHID DERHAB[3], MOHAMED BELAOUED[1,4], WALEED HALBOOB[3], JALAL AL-MUHTADI[3,5], ABDELGHANI BOURAS[6]

[1]LICUS Laboratory, Department of Computer Science, Université 20 Août 1955-Skikda, Skikda 21000, Algeria
[2]LIRE Laboratory, Abdelhamid Mehri-Constantine 2 University, Constantine 25000, Algeria
[3]Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh 11653, Saudi Arabia
[4]CReSTIC, University of Reims Champagne Ardenne, Reims, 51100 France
[5]College of Computer and Information Sciences, King Saud University, Riyadh 11653, Saudi Arabia
[6]Department of Industrial Engineering, College of Engineering, Alfaisal University,Riyadh 11533, Saudi Arabia

Corresponding authors: Abdelouahid Derhab (abderhab@ksu.edu.sa) and Waleed Halboob (wMohammed.c@ksu.edu.sa)

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no (RG-1441-531).

ABSTRACT The rapid development of different social media and content-sharing platforms has been largely exploited to spread misinformation and fake news that make people believing in harmful stories, which allow to influence public opinion, and could cause panic and chaos among population. Thus, fake news detection has become an important research topic, aiming at flagging a specific content as fake or legitimate. The fake news detection solutions can be divided into three main categories: content-based, social context-based, and knowledge-based approaches. In this paper, we propose a novel hybrid fake news detection system that combines linguistic and knowledge-based approaches and inherits their advantages, by employing two different sets of features: (1) linguistic features (i.e., title, number of words, reading ease, lexical diversity,and sentiment), and (2) a novel set of knowledge-based features, called *fact-verification* features that comprise three types of information namely, (i) *reputation of the website* where the news is published, (ii) *coverage*, i.e., number of sources that published the news, and (iii) *fact-check*, i.e., opinion of well-known fact-checking websites about the news, i.e., true or false. The proposed system only employs eight features, which is less than most of the state-of-the-art approaches. Also, the evaluation results on a fake news dataset show that the proposed system employing both types of features can reach an accuracy of 94.4%, which is better compared to that obtained from separately employing linguistic features (i.e., accuracy=89.4%) and fact-verification features (i.e., accuracy=81.2%).

INDEX TERMS Social media, Fake news detection, Linguistic analysis, Knowledge analysis, Fact-checking website.

I. INTRODUCTION

SOCIAL media are taking an increasing part in our professional and personal lives [21]. More and more people tend to search and consume news via social media rather than traditional media outlets. It has become common that important news are first broadcasted on social networks before being released by traditional media such as television or radio. Due to the massive propagation of news on social networks, users rarely check the accuracy of the information they share. It is therefore common to see false and manipulated information that are circulating on social media such as hoaxes, rumors [10], urban legends, and fake news [5], [5], [23], [41], [42], [82]. Moreover, it is difficult to stop the spreading of fake news when it is already shared many times and at large-scale [48]. This massive dissemination of false information [47], [61] could cause a serious negative impact on individuals and society. First, fake news could negatively influence the public opinion. Second, fake news change the way people interpret and react to real news. For example, some fake news could make people suspicious, and affect their ability to discern real news from fake news. In the literature, many approaches have been proposed for fake news detection. Early approaches were mainly based on linguistic-based techniques, which rely on language usage and its analysis to predict deception [3], [57], [60], [76]. The goal of these approaches is to look for instances of leakage found in the content of a text at different levels (i.e., words, sentences, characters, and documents levels). These approaches implement different methods such as: data representation, deep syntax, sentiment, and semantic analyses [17], [72]. In data representation methods, each word is considered as a single unit, and individual words are aggregated and analyzed to reveal linguistic cues of deception. In deep syntax methods, the sentences are converted into a set of rewritten rules (i.e., parse tree) in order to describe the syntax structure [12]. The semantic analysis determines the truthfulness of authors, which describes the degree of compatibility of personal experience compared to the content derived from a collection of analogous data. Finally, the sentiment analysis focuses on the extraction of opinion, which involves examining written texts about people's attitudes, sentiments, and evaluations using analytical techniques. Recent research has shown that linguistic-based techniques alone are not sufficient to reach a high detection accuracy, which generally does not exceed 80% [36], [37], [55], [57].

The knowledge-based approach is the most straightforward way to detect fake news, which allows to check the truthfulness of the statements claimed in news content [64]. Knowledge-based approaches [54] use external sources to verify if the news is fake or real and identify it before it spreads. This approach is divided into two distinct techniques [20] manual fact-checking, and automated fact-checking.

The manual fact-checking can be further divided into (a) crowd-sourced fact-checking, which is based on a large population of regular individuals acting as fact-checkers (i.e., collective intelligence), and (b) expert-based fact-checking, which is based on experts' judgments in the field (i.e., fact-checkers) to verify the content of the given news item [86]. Expert-based fact-checking is often performed by a small group of highly credible fact-checkers, which could lead to very accurate results. However, they require continuous and manual updates, and cannot perform automatic learning.

Through consultations and extraction of data from different sources, automated fact-checking aims at automatically verifying claims. Then, a classification based on the stance and strength of reputable sources regarding the claim is assigned [16]. Despite this technique is still in progress, it is very promising.

The major challenges that hinder the efficiency of the existing fake news detection solutions are related to the highly versatile nature of deceptive information. Indeed, it is very difficult to obtain a generalized dataset for fake news detection. Thus, it is very difficult to extract relevant features that can well represent and allow to detect fake news in various domains. In addition, it is also very challenging to detect fake news of a newly emerged event due to the limited information and knowledge regarding this event. For this reason, the use of one single technique for detecting fake content in news media will not able to reach the required level of efficiency.

In this paper, we propose a hybrid fake news detection system that takes advantage of both linguistic-based and knowledge-based approaches. The proposed system uses five different linguistic features, which are the title of the news, the number of words composing the news, its reading ease, its lexical diversity, and the dominant sentiment about the news. The system also employs three different knowledge-based features namely reputation, fact-check, and coverage. The system also implements four (04) different machine learning algorithms namely Random Forest (RF), Logistic Regression (LR), Additional Trees Discriminant (ATD), and XG-Boost. The earlier mentioned learning algorithms are trained and tested using different combinations of the aforementioned features, and the most performing classifier is selected. To the best of our knowledge, our work is the first that proposes this hybridization in the context of fake news detection. Specifically, the main contributions of the paper are the following:

- We propose a hybrid linguistic and knowledge-based fake news detection system that combines (1) linguistic features (i.e., title, number of words, reading ease, lexical diversity and sentiment), and (2) a novel set of knowledge-based features, called *fact-verification* features.
- The proposed fact-verification features allow to determine the truthfulness of a news trough the assessment of the reputation of the source (i.e., the website from which the information is obtained), and credibility to check if other fact-checking websites have already given their opinion about the news whether it is true or false.
- The proposed system only employs eight features, which is less than most of the state-of-the-art approaches.
- The evaluation results show that the proposed combination of features records more than 94% accuracy for fake news detection, and allows an increase of more than 7% compared to linguistic-based features.

The rest of the paper is organized as follows: Section II presents related work on fake news detection. In section III, we describe our proposed fake news detection system. Section IV describes the implementation of the proposed system. The evaluation results are presented in Section V, and discussed in Section VI. Finally, Section VII concludes the paper and highlights its key perspectives.

II. RELATED WORK

In this section, we provide a literature review of existing fake news detection solutions. As shown in Figure 1, fake news detection approaches can be divided into three categories namely, linguistic-based, social context-based, and knowledge-based. In the figure, some selected approaches from the literature are shown under each category, considered to be the most relevant ones in the last fifteen (15) years.

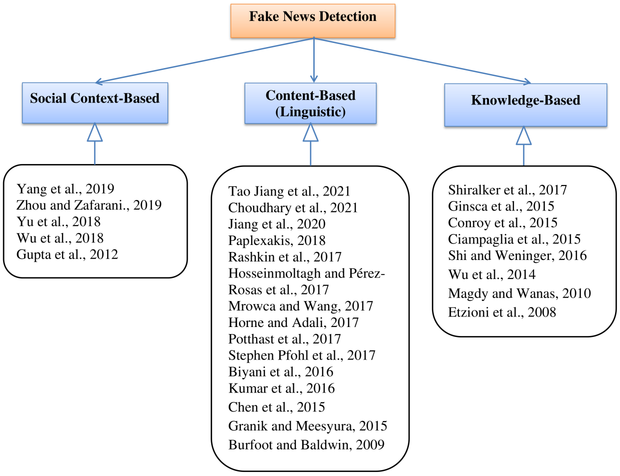


FIGURE 1. Taxonomy of fake news detection approaches

A. LINGUISTIC-BASED ANALYSIS

Linguistic analysis refers to the typical and the accurate examination of natural language. This approach [1], [19], [26], [83], [84] extracts valuable data from the news content, and examines the associated language patterns, meanings and structures of the news. As explained in [59], linguistic analysis mainly aims to identify the language competence of the news creator by the cognition of language formats and finding out the writing patterns. Data representation, deep syntax, semantic analysis and sentiment analysis are the main used techniques in linguistic analysis [17].

Many studies were conducted to examine the unique linguistic styles in clickbait articles such as in Biyani et al. [6]. Chen et al. [13] examined potential methods for the automatic detection of clickbait, which aim to find out both textual and non textual clickbait hints among images and users' behaviors.

Kumar et al. [43] proposed an approach for the identification of hoax documents in Wikipedia. Similar work [7], [62] detected different types of fake news based on the stance of headlines and by considering their corresponding article bodies. This approach could be applied on clickbait detection scenarios, and could be generalized to fake news detection.

Granik and Meesyura [28] introduced a simple and straightforward method for fake news detection based on machine learning techniques. The proposed approach used naive Bayes classifier and achieved good results considering the simplicity of their model.

Burfoot and Baldwin [9] proposed an approach for the automatic detection of satirical news. The proposed approach relies on Support Vector Machine (SVM) classifier, which is trained on lexical and semantic features (i.e., Headline, Profanity, Slang and validity). By applying Bi-normal separation feature scaling (BSN), a precision of 0.958 is achieved.

Rashkin et al. [58] studied and compared the language used by real news and that used by satire, hoax, and propaganda. They aimed to find linguistic features characterizing fake content.

Hosseinmoltagh and Paplexakis [37] investigated the problem of identifying the different types of fake news with high accuracy. To this end, they proposed a tensor model that captures the relations between the articles and terms, as well as the spatial relations among terms. In order to build high-coherent fake news clusters (i.e, clusters of similar types of fake news), the authors further proposed an ensemble method that combines the results of of multiple tensor decompositions.

To build a fake news detection model, Pérez-Rosas et al. [55] combined morphological, syntactic, understandability, psychological, and n-grams patterns. The authors observed that authentic news in newspaper and entertainment magazines are likely to begin with the first-person pronouns, besides containing positive feeling words. However, fake contents commonly utilize the second-person pronouns and negative feeling words, and focus on the present actions.

Mrowca and Wang [52] applied different deep neural network models on the Fake News Challenge (FNC)[1] dataset to solve the stance detection problem. The goal was to determine if there is any relationship between the headline and the news article. The results show that Word Embeddings Conditioned Bidirectional Long Short-Term Memory (LSTM) achieves the best classification accuracy.

To distinguish between satire, fake, and original news articles, Horne and Adali [36] and Potthast et al. [57] proposed models that consider complexity, stylistic and psychological aspects. In [36], the results show that fake news are similar to satire news with respect to content. On the other hand, it is very easy to distinguish between real news and fake news. Potthast et al. [57] constructed two categorizations models for the purpose of distinguishing between satire, fake, mainstream, and hypopartisan news articles. The first model is called topic-based and the second one is known as style-based. The two categorization models are proved to be efficient at differentiating between hyperpartisan news and mainstream news.

Pfohl et al. [56] addressed the stance detection problem, which aims at identifying the relationship between the headline and the body text of the news article. They applied four neural models namely, Bag of Words (BoW), LSTM, LSTM with attention, and conditional encoding LSTM with attention on Fake News Challenge dataset. The results show that the attention models achieve better results than BoW and LSTM with respect to F1 Score.

Jiang et al. [38] proposed a fake news detection system, which applied BiLSTM and text embedding Glove on linguistic-based features. The proposed approach was able to achieve an accuracy of 99.82%. In a more recent work [39],the same authors introduced a new stacking method for fake news detection that consists of training nine (09) different machine learning and deep learning algorithms on a set of linguistic features that were extracted using Term

[1] Available at : http://www.fakenewschallenge.org/

Frequency- Inverse Document Frequency (TF-IDF) feature extraction.

Choudhary et al. [14] introduced a fake news detection system based on linguistic features. The authors extracted four different linguistic features, namely, syntactic, grammatical, sentimental, and readability features. In order to classify news into either fake or real, the authors used a sequential neural network (SNN), which was trained on different combinations of features, and thus building various linguistic-features-based SNN models. Experimental results showed that the classification model, which was trained on combined features, achieved the highest performances with 86% of accuracy using a dataset composed of 250 news.

Some metaheuristics algorithms have been proposed to deal with the fake news detection issue. In [53], two metaheuristic algorithms, i.e, salp swarm optimization (SSO) and grey wolf optimization (GWO) were proposed. In [65], the authors proposed a linguistic-based fake news detection system that applies the Extreme Gradient Boosting Tree (xgbTree) algorithm, which is optimized by the Whale Optimization Algorithm (WOA). Al-Ahmad et al. [2] proposed an approach that aims at reducing the number of symmetrical features that exist in news, and particularly in COVID-19 pandemic news. To do so, they implemented different evolutionary classifications such as Salp swarm algorithm (SSA), particle swarm optimization (PSO), and genetic algorithm (GA). Zivkovic et al. [87] also focused on detecting misinformation related to COVID- 19 pandemic. They proposed an arithmetic optimization algorithm (AOA) as a a wrapper feature selection to reduce the number of features, and combined it with KNN classifier.

Despite the fact that linguistic-based solutions can effectively detect fake news, some research works [17], [44] have shown that relying only on linguistic analysis, is not suitable for designing robust fake news detection systems. We also agree with this opinion, knowing that the effectiveness of a linguistic-based solution is closely tied to data veracity, which is difficult to ensure.

B. KNOWLEDGE-BASED ANALYSIS

Knowledge-based analysis aims to complement the content–based approaches such as the linguistic ones, by checking the existing body of human knowledge to estimate the likelihood of new statements to be false. The method [17] allows to collect and compare a large number of common and connected statements from different networks like metatags and social network behavior to compute the probability that the content is fake. According to [68], [74], knowledge–based analysis and particularly fact checking, aims at using external sources to check the truthfulness of claims in news contents.

Magdy and Wanas [46] measured the support for each fact of the document using web search. The measured supports are accumulated to compute the support of the document. According to Ginsca et al. [27], the technique in [46] has to take into consideration the different aspects of web information credibility such as: quality, expertise, trustworthiness, and reliability. Etzioni et al. [24] proposed an approach for fake news detection based on knowledge analysis, which consists of matching the claims extracted from the web with the analyzed news story.

Some existing solutions rely on ontologies in order to model fake news domain knowledge, which can be then used to distinguish fake from real news content. For instance, in [29], ontology reasoning and natural language processing (NLP) have been combined in order to detect deceptive information about COVID-19. The major challenge that faces this specific category of news, is the lack of scientific knowledge related to the disease. For this purpose, the proposed approach applies Description Logics semantic reasoning and NLP in order to identify inconsistencies between trusted and non-trusted medical sources. For instance, the study demonstrated that trusted news are written in a formal language, unlike non-trusted ones, which are written in a less formal way. Similarly, Mazepa et al. [51] suggested an ontology for fake news detection on social networks, and Hamilton [31] focused on identifying propaganda techniques in news articles.

Wu et al. [79] proposed a fact-checking framework, which applies different perturbations on the claims and checking the corresponding results.

Shi and Weninger [66] formulated the fact checking problem as a link-prediction algorithm in a knowledge graph, Ciampaglia et al. [15] used the shortest path between concepts in a knowledge graph. The approaches in [15], [66] are inappropriate for new claims due to the lack of corresponding entries in knowledge bases.

Ciampaglia et al. [15] addressed fact-checking as a network problem, through the use of Wikipedia infoboxes to draw out truths in an organized manner. They suggested a measurement to evaluate the truthfulness of a statement by studying path lengths between concepts and the specificity of the terms of the claim in the Wikipedia knowledge graph.

Shiralker et al. [67] proposed a fact-checking algorithm called Relational Knowledge Linker, which converts the knowledge network to a smooth–continuous network and examines a claim on the single shortest and semantically connected path in the knowledge graph.

Although knowledge-based approaches can achieve good results, they are inappropriate for new claims without corresponding entries in a knowledge base. Thus, relying only on knowledge-based analysis to build a fake news detection system is not recommended.

C. SOCIAL CONTEXT ANALYSIS

The social context-based approaches [30], [45], [78], [80], [85] typically analyze the spreading patterns and the diffusion on social networks to distinguish misleading substance. Yang et al. [80] proposed an unsupervised approach for fake news detection on social media. The authors investigated the veracity of news and credibility of users, and utilized a probabilistic graphical model to capture the complete generative spectrum. They evaluated the model on two different datasets (i.e., LIAR and BuzzFeedNews), and obtained an accuracy of 75.9% and 67.9% respectively.

Zhou and Zafara [85] proposed a network-based pattern-driven approach for fake news detection in social networks. The main idea behind this work is to focus on the credibility of the news source, covering both the sources that create and publish the news, as well as the sources that spread the news. The method was evaluated on PolitiFact and BuzzFeed datasets showing good performance compared to the state of the art, with an accuracy of 93.30%.

Wu and Huan [78] proposed an approach for social media news classifying using diffusion traces in social networks. They first inferred embeddings of social media users with social network structures to classify news items. To this end, they utilized a new Long Short-Term Memory networks (LSTM-RNN) model to represent and classify the propagation pathways of a message. They evaluated the proposed model on real-world datasets, and the experimental results demonstrated its effectiveness on the task of fake news detection and categorization.

Yu et al. [45] used a combination of recurrent and convolutional neural networks to model news diffusion pathways as multivariate time series, where each tuple of a news story is a numerical vector representing characteristics of a user who engaged in spreading the news. The method was evaluated on three real-world datasets and experimental results showed that the proposed model was able to effectively identify fake news content with an accuracy of 92.3%.

Gupta et al. [30] proposed a PageRank-like credibility propagation algorithm on a multi-typed network by encoding users' credibility and tweets' implications. Further, they enhanced the basic trust analysis by updating event credibility scores and exploited event graph-based optimization to assign similar scores to similar events. They evaluated the model on two tweet feed datasets, and the proposed approach achieved an accuracy of 86%.

In general, context-based approaches achieve better performance compared to both linguistic and knowledge-based ones. However, they work only a posteriori, disregarding the actual news content and requiring large amounts of data.

D. COMPARISON WITH RELATED WORK

As previously discussed, the existing fake news detection solutions are either linguistic-based, knowledge-based, or social context-based. Considering the limitations of the aforementioned categories, it would be a good idea to investigate combining two different categories in order to overcome their respective limitations. In the literature, there are only few hybrid approaches, that only considered combining linguistic and social context analyses [10], [63], [69], [75]. To the best of our knowledge, there is no hybrid approach that considers both linguistic and knowledge-based features. Differently from related work, we propose in our work a hybrid approach, which combines linguistic-based and knowledge-based analyses to build a more robust and accurate fake news detection system.

III. PROPOSED FAKE NEWS DETECTION SYSTEM

The proposed fake news detection system consists of two phases, namely training and testing. Both phases include a preprocessing task, which consists of cleaning and preparing the training and testing datsets of real and fake news. In the training phase, the feature extracting task extracts a set of relevant features from the training dataset, which are then fed to several machine learning algorithms to build a fake news detection model. In the testing phase, the detection model is applied on test data to decide whether the provided news articles are real or fake. Figure 2 presents the overall architecture of the proposed fake news detection system.

A. PRE-PROCESSING

Before extracting the various features and analyzing the news content, we need to conduct a pre-processing task. In this work, we apply the following text processing methods:

- Tokenization: It consists of splitting a news content into a set of individual words.
- Stopwords removal: It consists of removing the most commonly used words (e.g., the, and, is), which have no effect on the classification.
- Stemming: It consists of reducing a word either to its base form by removing suffixes and prefixes or to its root form, also known as a lemma.
- Cleaning: It consists of removing URLs, punctuation, .etc.

B. FEATURE EXTRACTION

In machine learning, features are usually numerical, but structural features such as strings and graphs can also be used. In the context of our work, features represent different properties of the news article, such as its title, the number of words, sentiment, etc.

In our work, we use a set of linguistic features, which have been considered by [33], [36], [55], [70] as the most relevant ones for distinguishing between real and fake news. These features are: title, number of words, reading ease, lexical diversity, and sentiment.

In addition to the linguistic features, we propose our own set of features, called fact-verification, which are: Fact Check (FC), Reputation (Rep), and Coverage (CV).

By combining linguistic and fact-verification features, we expect to obtain a better detection accuracy, since we leverage the benefits of both categories of features. Table 1 defines the set of feature, which are used by the proposed detection system.

1) Fact-verification features

We list the used fact-verification features and their corresponding extraction methods:

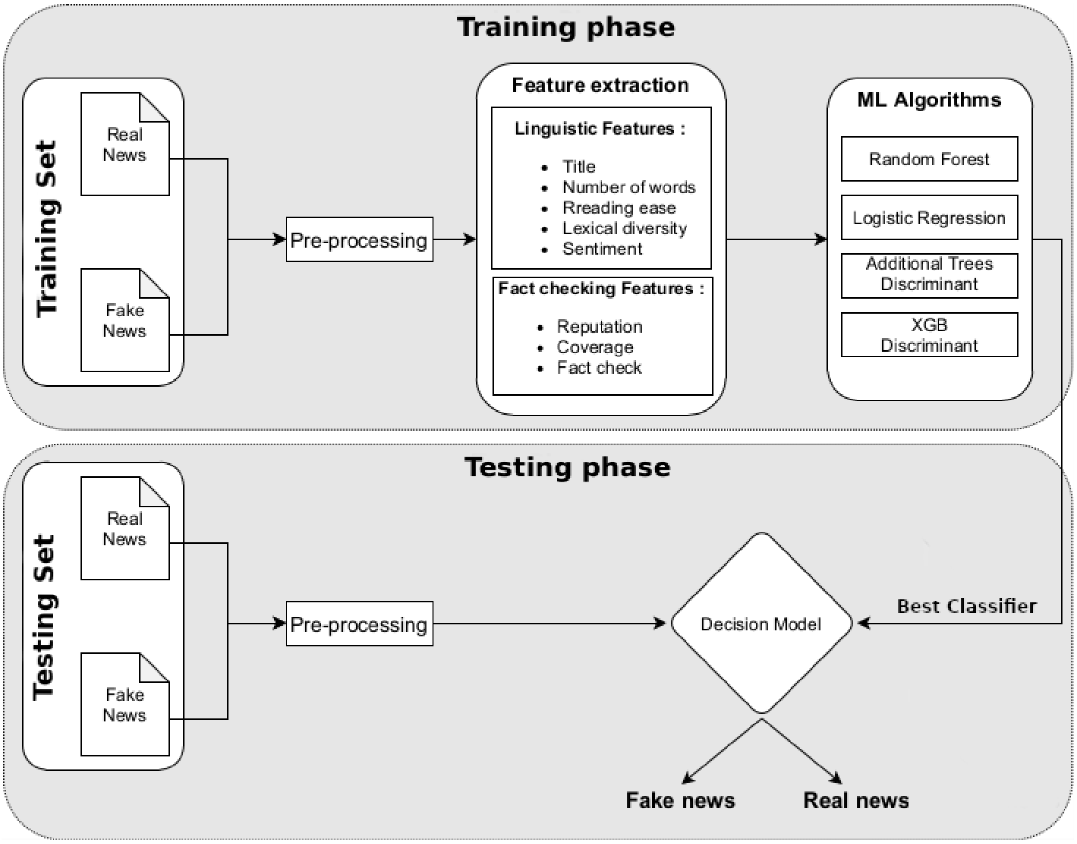


FIGURE 2. Architecture of the proposed fake news detection system

TABLE 1. List of Linguistic and fact-verification features used by the proposed detection system

Features	Type	Description
Title	Linguistic	The title of the news article
Reputation	Fact-verification	It allows to rate the website that published the article. A website can have good, bad, or no reputation if it is not known.
N of word	Linguistic	It represents the number of words
Reading ease	Linguistic	It represents the reading difficulty of each article, as shown in Equation 1. The higher the score, the easier the article is to read.
Lexical diversity	Linguistic	It is an aspect of ’lexical richness’ and refers to the relationship between different types of words and the total number of words.
Sentiment	Linguistic	It indicates the dominant sentiment that is present in the article. It can be either positive, negative or neutral.
Fact check	Fact-verification	It is used to check if other fact checking sites have already given their opinion about the article, i.e., whether it is true or false.
Coverage	Fact-verification	This feature answers the question: How many credible sources published the same information ?

- **Fact check:** It is extracted using fact-checking websites (Snopes[2] and Google Fact Check Explorer[3]).
- **Reputation:** It is obtained using two different tools: Decodex of the French newspaper Le Monde[4] and Media Bias Fact Check[5], a famous American fact checking site.
- **Coverage:** It is extracted using the google search engine[6].

2) Linguistic features

We list the following linguistic features and their corresponding extraction tools:

- **Title:** It is obtained using two web-based software: Zyte[7] and BuzzStream[8].

[2] Available at : https://www.snopes.com/
[3] Available at: https://toolbox.google.com/factcheck/explorer
[4] Available at: https://www.lemonde.fr/verification
[5] Available at: https://mediabiasfactcheck.com/
[6] Available at: https://www.google.com/
[7] Available at : https://www.zyte.com/data-extraction/
[8] Available at: http://tools.buzzstream.com/meta-tag-extractor

- **N of words:** It is obtained using different tools such as WordCounter[9].
- **Reading ease:** It computes the reading ease of a text using the Readability Formulas tool, which provides the Flesch Reading Ease score, as shown in Equation 1.

$$RE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW) \quad (1)$$

where :

- RE : Readability Ease.
- ASL : Average Sentence Length (i.e., the number of words divided by the number of sentences).
- ASW : Average number of syllables per word (i.e., the number of syllables divided by the number of words).

The Flesch Reading Ease Formula provides a score between 1 and 100, and 100 represents the best readability score. According to its reading ease score, a text is classified as follows:

  - **90-100 :** Very easy.
  - **80-89 :** Easy.
  - **70-79 :** Fairly easy.
  - **60-69 :** Standard.
  - **50-59 :** Fairly difficult.
  - **30-49 :** Difficult.
  - **0-29 :** Very confusing.

- **Lexical diversity:** It can be computed using a variety of measures by considering the following variables:
- V: is the number of types of tokens.
- N: is the total number of tokens.
- $f_v(i, N)$: is the numbers of types occurring i times in a sample of length N.

The lexical diversity measures are the following:

  - *Type-Token Ratio* [35] (Equation 2), denoted by:

$$TTR = \frac{V}{N} \quad (2)$$

  - *Log Type-Token Ratio* (or *Herdan's C*) [73] (Equation 3), denoted by:

$$C = \frac{\log V}{\log N} \quad (3)$$

  - *Guiraud's Root TTR* [73] (Equation 4), denoted by:

$$R = \frac{V}{\sqrt{N}} \quad (4)$$

  - *Carroll's Corrected TTR* [49] (Equation 5), denoted by:

$$CTTR = \frac{V}{\sqrt{2N}} \quad (5)$$

  - *Dugast's Uber Index* [73] (Equation 6), denoted by:

$$U = \frac{(\log N)^2}{\log N - \log V} \quad (6)$$

  - *Summer's index* [49] (Equation 7), denoted by:

$$S = \frac{\log(\log V)}{\log(\log N)} \quad (7)$$

  - *Yule's K* [81] (Equation 8), denoted by:

$$K = 10^4 \times [-\frac{1}{N} + \sum_{i=1}^{v} f_v(i, N)(\frac{i}{N})^2] \quad (8)$$

  - *Yule's I* [81] (Equation 9), denoted by:

$$I = \frac{V^2}{M_2 - V} \quad (9)$$

$$where \quad M_2 = \sum_{i=1}^{v} i^2 \times f_v(i, N)$$

  - *Simpson's D* [71] (Equation 10), denoted by:

$$D = \sum_{i=1}^{v} f_v(i, N) \frac{i}{N} \frac{i-1}{N-1} \quad (10)$$

  - *Herdan's V_m* [34] (Equation 11), denoted by:

$$V_m = \sqrt{\sum_{i=1}^{v} f_v(i, N)(i/N)^2 - \frac{i}{V}} \quad (11)$$

  - *Maas' indices* [50] (Equation 12), denoted by:

$$Maas = (a, \log V_0, \log_e V_0) \quad (12)$$

$$where \quad a^2 = \frac{\log N - \log V}{\log N^2}$$

$$and \quad logV_0 = \frac{\log V}{\sqrt{1 - \frac{\log V^2}{\log N}}}$$

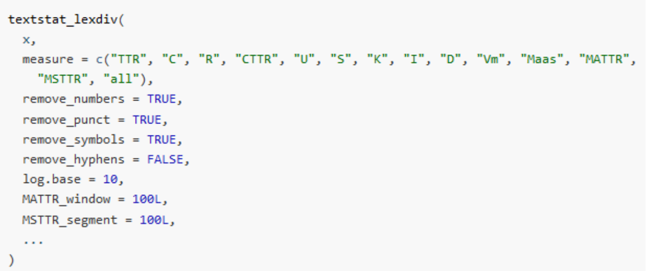
  - *Moving-Average Type-Token Ratio* [18], denoted by $MATTR$. It is computed by moving a fixed size window through the text, compute the TTR of every window, and average the obtained $TTRs$.
  - *Mean Segmental Type-Token Ratio* [40], denoted by $MSTTR$. It is computed by dividing the text into segments, compute the TTR of each segment, and average the obtained $TTRs$.

The above equations are computed using Quanteda tool [4], as shown in Figure 3. In our case, we define our lexical diversity as the sum of all values obtained from Equations 2 to 12, $MATTR$, and $MSTTR$.

- **Sentiment:** We used the SEO Scout's analysis tool[10] to get this feature, as it can effectively and rapidity estimate the dominant sentiment in the news.

[9] Available at: https://wordcounter.net/

[10] Available at: https://seoscout.com



```
textstat_lexdiv(
  x,
  measure = c("TTR", "C", "R", "CTTR", "U", "S", "K", "I", "D", "Vm", "Maas", "MATTR",
    "MSTTR", "all"),
  remove_numbers = TRUE,
  remove_punct = TRUE,
  remove_symbols = TRUE,
  remove_hyphens = FALSE,
  log.base = 10,
  MATTR_window = 100L,
  MSTTR_segment = 100L,
  ...
)
```

FIGURE 3. Lexical diversity measures in quanteda tool

C. TRAINING PHASE

In the training phase, we leverage AutoAI experiment of IBM Watson Studio[11] to select the best learning algorithm among a set of candidate ones. The best algorithm is the one that offers the best match for training data. To this end, we employ four candidate classification algorithms namely, Random Forest (RF), Logistic Regression (LR), Additional Trees Discriminant (ATD), and eXtreme Gradient Boosting (XGBoost). The best model, which is selected by AutoAI experiment is Random Forest. To select the best model, AutoAI initially applies the candidate algorithms on small subsets of the dataset, and ranks them. Then, it repetitively increases the size of subsets and executes the candidate algorithms until the best algorithm is found.

AutoAI performs hyper-parameter optimization of the best selected algorithm by applying an optimization algorithm that allows fast convergence to a good solution, and generating the best model. Note that AutoAI automatically selects the hyper-parameters of each machine learning algorithm. Thus, we did not apply any parameter tuning.

1) Random Forest Classifier (RF)

Random forest [8] is a supervised classification algorithm that consists of a set decision trees, which are merged together for better performance in terms of accuracy. Each tree in the random forest produces a class prediction, and the class with the majority of predictions becomes the model's prediction. The RF algorithm, introduced by Dietterich et al. [22], describes the steps of constructing the decision tress as follows :

1) Take L instances of M attributes from the training set.
2) $m < M$, is the number of parameters in the training set that determines the next selected attribute at each node.
3) For each training sample, a tree is constructed with replacement.
4) Arbitrarily select m attributes for each node of the tree.
5) Compute the best split using m training set's attributes.
6) Grow each tree without pruning.

2) Logistic Regression (LR)

Logistic regression [77] is a linear algorithm used for binary classification problems. Linear and logistic regressions are very similar, but the main difference is that linear regression generates a continuous output, and logistic regression generates a discrete one. This algorithm allows the description of the data and the level of strength in the relationship between a dependant binary variable and the associated independent variables. In other words, Logistic Regression predicts a categorical dependent variable representing the target class based a given set of independent variables representing the features' set.

3) Additional Trees Discriminant (ATD)

Decision Trees are supervised machine learning algorithms where data is segmented according to a specific parameter. The objective of this algorithm is to build a training model that is used to predict the class of the target variable, which is used by the decision tree to solve the classification problem. Discriminant analysis creates a predictive model that determines to which group the class belongs. The model is composed of a discriminant function based on linear combinations of the variables used as predictors, i.e., offering the best discrimination between the groups.

4) XGBoost (Extreme gradient boosting)

XGBoost [11] is a decision-tree-based ensemble machine learning algorithm that implements the Gradient Boosting method. Gradient boosting is a supervised learning algorithm aims at providing accurate prediction of a target variable by combining the estimates from other models. XGBoost offers parallel construction of trees, as well as an optimization step for each attached tree (i.e., boosting). Moreover, XGBoost employs regularization that helps avoiding overfitting when training the model. All these characteristics make XGBoost one of the widely used machine learning algorithms that allows solving various regression and classification problems in a fast and an accurate manner.

D. TESTING PHASE

In the testing phase, each news is pre-processed. Then, the extracted features are fed to the decision model, which is selected in the training phase. The decision model decides whether the news is real or fake.

IV. IMPLEMENTATION

In this section, we present the software and hardware tools, which are used to implement our system.

A. IMPLEMENTATION ENVIRONMENT

To implement our learning models, we use IBM Watson Machine Learning software[12]. This software can generate analysis models, which are trained on our dataset.

Watson Machine Learning offers a full range of tools and services to generate, train, and deploy machine learning models.

[11] Available at: https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/autoai-overview.html

[12] Available at : https://eu-gb.dataplatform.cloud.ibm.com/login?context=cpdaas

The following tools are available with the Watson Machine Learning service:

- The AutoAI experimentation generator, which automatically processes structured data to generate model pipelines. The best performing pipelines can be saved as machine learning models and deployed for evaluation and the best algorithm is determined by the model selection and optimization during AutoAI training.
- Notebooks that provide an interactive programming environment for working with data, testing models, and obtaining rapid prototyping.
- Deep learning experiments, which automate the execution of hundreds of training runs while tracking and storing results.
- Tools to view and manage model deployments.

The hardware and software settings, which are used to implement our system, are depicted in Table 2.

TABLE 2. Experimental environment of the proposed detection system

Component	Settings
	Hardware Settings
CPU	Intel(R) Core(TM) i5-4200U CPU @ 1.60 GHz 2.30 GHz
RAM	8 GB
	Software Settings
OS	Windows 10(32-bit)
Watson Machine Learning Service instance	Machine Learning-3e
Notebook	2 Executors: 1 vCPU and 4 GB RAM, Driver: 1 vCPU and 4 GB RAM
Programming Language	Python 3.7 with Spark

B. DATASET

In our experiments, we use the Buzzfeed Political News data set[13] that have been proposed by Horne and Adali [36]. This dataset contains two categories of news, namely, fake and real, and it has been gathered from a Buzzfeed's 2016 article[14] on fake news election that have been spread on Facebook.

V. EVALUATION

In this section, we present the evaluation methodology, the evaluation metrics, and the performance results of our system in terms of accuracy, recall, and F1-score.

A. EVALUATION METHODOLOGY

Our experimental process consists of training and testing our system using only linguistic features, then using only fact-verification features. Finally, we train and test the proposed system using various combinations of linguistic and fact-verification features. In order to assess the effectiveness of the proposed fake news detection system, we used the holdout cross-validation technique, which consists in dividing the dataset into two subsets: training and testing. In our case, we choose to use 85% of the data for training, and 15% of the data for testing.

B. EVALUATION METRICS

To evaluate the performance of our system, we used several evaluation metrics, which are: Accuracy (ACC), Recall (REC), and F1-score ($F1$) [32].

Accuracy (ACC): It is a measure of the proportion of correct predictions of the model, and is defined as the number of true predictions divided by the total number of analyzed items. Formally:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Recall (REC): It corresponds to the ratio of the number of correctly classified positive items to the number of actual positive items. Formally:

$$REC = \frac{TP}{TP + FN} \tag{14}$$

F1-score ($F1$): It relates precision (PRE) and recall (REC) metrics to obtain a quality measure that balances the relative importance of these two metrics. Formally:

$$F1 = \frac{2 \times (PRE \times REC)}{PRE + REC} \tag{15}$$

where:

$$PRE = \frac{TP}{TP + FP} \tag{16}$$

- TP : It is the number of accurately classified fake news (true positives).
- FP : It is the number of incorrectly classified real news (false positives).
- TN : It is the number of accurately classified real news (true negatives).
- FN : It is the number of inaccurately classified fake news (false negatives).

C. EVALUATION RESULTS

Table 3 presents the performance results of our system. We can observe that the linguistic-based features give an accuracy of 89.4% under ATD and XGBoost algorithms. On the other hand, fact-verification features gives an accuracy of 81.2% under ATD and RF algorithms. We can observe that the combination of the linguistic features and fact-verification features considerably increases the accuracy, especially when the eight (08) features (i.e., title, number of words, reading ease, lexical diversity, sentiment, Fact check, Coverage, and Rep) are employed together. Indeed, we obtain the highest accuracy, i.e., 94.40%, which represents an improvement of respectively, 5%, and 13% compared to the highest accuracy obtained when linguistic features, and fact-verification features are employed separately.

[13] Available at: https://github.com/rpitrust/fakenewsdata1

[14] Available at: https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

TABLE 3. Experimental Results of the proposed detection system under different sets of features

Features	Learning Algorithm	F1 (%)	REC(%)	ACC (%)
Fact check	LR	69.40	54.40	75.30
	ATD	69.40	54.40	75.30
	XGBoost	70.40	54.40	76.50
	RF	70.40	54.40	76.50
Linguistic	LR	79.60	84.10	77.60
	ATD	89.60	88.90	89.40
	XGBoost	90.00	91.10	89.40
	RF	88.30	93.20	87.00
Linguistic + Fact check	LR	84.30	91.10	82.30
	ATD	88.90	84.30	89.50
	XGBoost	88.40	86.30	88.20
	RF	91.20	93.30	90.60
Fact-verification (Fact check + Coverage + Rep)	LR	76.7	63.70	80.00
	ATD	77.80	63.70	81.20
	XGBoost	75.10	61.40	78.80
	RF	77.80	63.70	81.20
Linguistic + Fact-verification (Fact check + Coverage + Rep)	LR	84.30	91.10	85.10
	ATD	89.10	84.10	88.40
	XGBoost	87.10	84.10	90.50
	RF	**94.90**	**97.90**	**94.40**

D. COMPARISON WITH OTHER APPROACHES

Actually, a fair comparison with previous approaches is not possible due to many reasons including the use of different methods (i.e., machine learning, deep learning, neural Networks, ...etc), as well as different datasets. However, we can only compare our work (Accuracy=94.40%) with Horne and Adali [36] (Accuracy=77%), as the same data set, i.e., Buzzfeed Political News Data, is used.

In Table 4 and Figure 4, we present the performances of our system and those of machine learning-based state-of-the-art approaches with respect to accuracy and the number of features. Zhou and Zafarani [85] employed linguistic and social features, and achieved a detection accuracy of 93.30%. Gupta et al. [30] and Yang et al. [80], which only used social-context features, obtained an accuracy of 86% and 75.90% respectively. Shu et al. [69] and Castillo et al. [10] achieved an accuracy of 87.80% and 89% respectively. Almeida et al. [19], Potthast et al. [57], Horne and Adali [36], Pérez-Rosas et al. [55], and Fairbanks et al. [25], which only used linguistic-based features, obtained an accuracy of 74.30%, 75%, 77%, 78%, and 88% respectively. Shakeel and Jai [64] that applied only knowledge-based features, reached an accuracy of 86%.

The efficiency of the detection approach, with respect to detection time, mainly depends on the used number of features. From Table 4, we can observe that our work only employs eight features, which is less compared to most of the state-of-the-art approaches. For instance, 10, 15, 17, 20, 26, 68, and 69 features are used by [69], [80], [57], [85] [55], [10], and [19] respectively. On the other hand, Horne and Adali [36], which uses the same dataset as our work, employs the lowest number of features (4 features). However, it records poor effectiveness (i.e., ACC=77%). Therefore, our system offers a good tradeoff between effectiveness and number of features.

VI. DISCUSSION

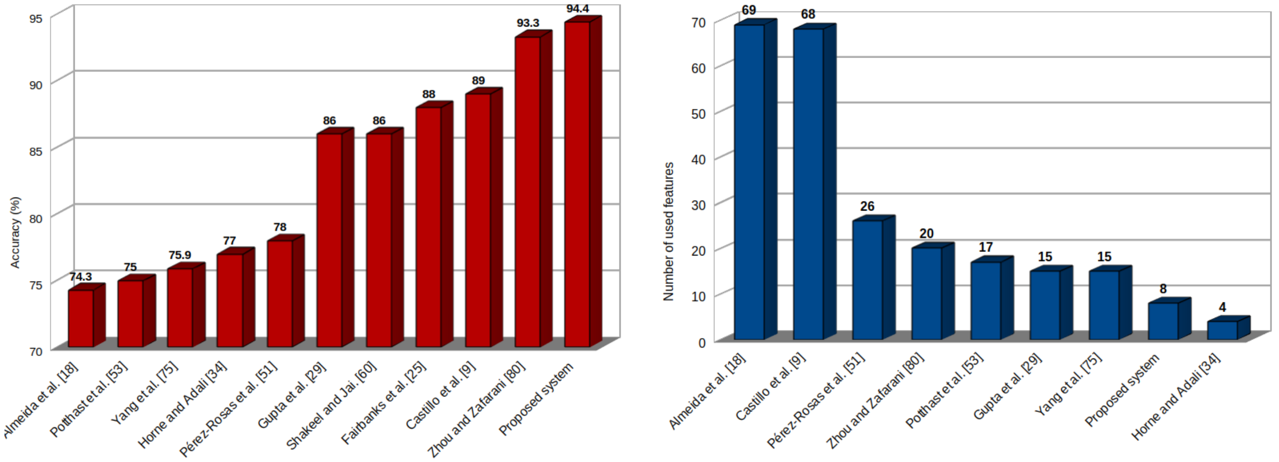
The evaluation results from Table 3 and Table 4 consolidate the claim that combining linguistic-based and knowledge-based features is an effective approach for fake news detection. Indeed, by employing linguistic features, our system reached an accuracy between 77.6% and 89.46%. The approaches that only employed linguistic-based features [19], [25], [36], [55], [57] recorded an accuracy between 74.3% and 88%. Specifically, Almeida et al. [19], Potthast et al. [57], Horne and Adali [36], Pérez-Rosas et al. [55], and Fairbanks et al. [25], which used linguistic-based features, obtained an accuracy of 74.30%, 75%, 77%, 78%, and 88% respectively. This shows that linguistic-based features are not sufficient for fake news detection and they provide poor accuracy performance.

On the other hand, we tested our system on two types of knowledge-based features. The first type only considers the Fact check feature, which contributed in reaching an accuracy between 75.3% and 76.5%. The second type considers the fact-verification features (i.e., Fact check, Coverage, and Reputation), which improved the accuracy to reach values between 78.8% and 81.2%. Shakeel and Jai [64], which only employed knowledge-based features, recorded an accuracy of 86%. The approaches that only employed social context features such as Yang et al. [80] and Gupta et al. [30] only recorded an accuracy of 75.96% and 86% respectively. These results indicate that knowledge-based features and the social-context features alone still cannot be a good choice for fake news detection.

By combining different types of features, a better detection accuracy can be obtained. For instance, Shu et al. [69], Castillo et al. [10], and Zhou and Zafarani [85], which com-

TABLE 4. Comparison with related work tested under different datasets

Work	Decision method	Dataset	Features	Number of Features	ACC
Pérez-Rosas et al. [55]	Support Vector Machine	Celebrity	Linguistic	26	78%
Fairbanks et al. [25]	Random Forest,Logistic Regression, Loopy Belief Propagation	Private	Linguistic	N/A	88%
Gupta et al. [30]	Support Vector Machine, Naive Bayes, Decision Trees	Twitter	Social Context	15	86%
Shu et al. [69]	Support Vector Machine	Fakenewsnet	Linguistic+Social Context	10	87.80%
Yang et al. [80]	Support Vector Machine, Naive Bayes	LIAR	Social Context	15	75.90%
Potthast et al. [57]	Random Forest	BuzzFeed-Webis	Linguistic	17	75%
Shakeel and Jai [64]	Support Vector Machine,Logistic Regression	DBpedia	Knowledge	N/A	86%
Zhou and Zafarani [85]	Random Forest, Support Vector Machine, Naive Bayes , Decision Trees	Fakenewsnet	Linguistic + Social Context	20	93.30%
Almeida et al. [19]	k-Nearest Neighbor, Support Vector Machine, Naive Bayes, Random Forest	Fakenewsnet	Linguistic	69	74.30%
Horne and Adali [36]	Support Vector Machine	BuzzFeedNews	Linguistic	4	77%
Castillo et al. [10]	Decision Tree	Twitter	Linguistic+Social Context	68	89%
Proposed system	**Random Forest**,Logistic Regression, Additional Trees Discriminant, XGBoost	BuzzFeedNews	Linguistic+fact-verification	8	**94.40%**



(a) Accuracy

(b) Number of features

FIGURE 4. Performance of fake news detection approaches

bined between linguistic-based features and social-context features, reached an accuracy of 87.7%, 89%, and 93.3% respectively.

As for the combination between linguistic-based and knowledge-based features, we tested our system under two combinations. The first combination considers the linguistic-based features and Fact check feature. This allowed to record an accuracy between 82.3% and 90.6%. The second combination considers the linguistic and the fact-verification features, i.e., after trying various combinations of these features, we find that the combination of all the eight features i.e., title, number of words, reading ease, lexical diversity, sentiment, Fact check, Coverage, and Reputation produces the best performance results. Indeed, this combination allows to reach an accuracy of 94.40% with RF algorithm. This represents a considerable leap in the effectiveness of the proposed system, since the obtained accuracy is 5% higher than that obtained using only linguistic features, and 13% higher than that obtained using fact verification ones.

VII. CONCLUSION

In this paper, we have proposed a novel hybrid fake news detection system that employs two types of features: linguistic and fact-verification features.

The proposed detection system employs only eight features, which less compared to the stat-of-the-art approaches.

It operates in two phases: training and testing. In the training phase, the detection system runs four machine learning algorithms, i.e., Logistic Regression (LR), Random Forest (RF), Additional Trees Discriminant, and XGBoost, in order to select the best classifier for the testing phase.

Evaluation results on the Buzzfeed Political News data set show that the proposed detection system achieves an accuracy of 94.4% under Random Forest. These results are better compared to those obtained from employing linguistic features, (i.e., Accuracy=89.4% under ATD and XGBoost), and fact-verification features (i.e., Accuracy=81.2% under ATD and Random Forest). The proposed system also employs eight features, which is less than most of the state-of-the-art approaches.

As future work, we aim at improving the accuracy of our detection system by investigating other discriminating features such as visual-based and style-based features. Moreover, we plan to further detect other types of false information such as biased/inaccurate news and misleading/ambiguous news.

REFERENCES

[1] Hadeer Ahmed, Issa Traore, and Sherif Saad. 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, pages 127–138. Springer, 2017.

[2] Bilal Al-Ahmad, Ala'M Al-Zoubi, Ruba Abu Khurma, and Ibrahim Aljarah. An evolutionary fake news detection method for covid-19 pandemic information. Symmetry, 13(6):1091, 2021.

[3] Sean Baird, Doug Sibley, and Yuxi Pan. Talos targets disinformation with fake news challenge victory. Fake News Challenge, 2017.

[4] Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. quanteda: An r package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30):774, 2018.

[5] Dan Berkowitz and David Asa Schwartz. Miley, cnn and the onion: When fake news becomes realer than real. Journalism Practice, 10(1):1–17, 2016.

[6] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. " 8 amazing secrets for getting more clicks": detecting clickbaits in news streams using article informality. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[7] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, pages 84–89, 2017.

[8] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[9] Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In Proceedings of the ACL-IJCNLP 2009 conference short papers, pages 161–164, 2009.

[10] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web, pages 675–684, 2011.

[11] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015.

[12] Yimin Chen, Nadia K Conroy, and Victoria L Rubin. News in an online world: The need for an "automatic crap detector". Proceedings of the Association for Information Science and Technology, 52(1):1–4, 2015.

[13] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: recognizing clickbait as" false news". In Proceedings of the 2015 ACM on workshop on multimodal deception detection, pages 15–19, 2015.

[14] Anshika Choudhary and Anuja Arora. Linguistic feature based learning model for fake news detection and classification. Expert Systems with Applications, 169:114171, 2021.

[15] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. PloS one, 10(6):e0128193, 2015.

[16] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. Computational journalism: A call to arms to database researchers. In CIDR, 2011.

[17] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. Proceedings of the association for information science and technology, 52(1):1–4, 2015.

[18] Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of quantitative linguistics, 17(2):94–100, 2010.

[19] Thais Gomes de Almeida. Liardetector: a linguistic-based approach for identifying fake news. 2019.

[20] Dylan de Beer and Machdel Matthee. Approaches to identify fake news: A systematic literature review. In International Conference on Integrated Science, pages 13–22. Springer, 2020.

[21] Abdelouahid Derhab, Rahaf Alawwad, Khawlah Dehwah, Noshina Tariq, Farrukh Aslam Khan, and Jalal Al-Muhtadi. Tweet-based bot detection using big data analytics. IEEE Access, 9:65988–66005, 2021.

[22] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2):139–157, 2000.

[23] Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann. Belittling the source: trustworthiness indicators to obfuscate fake news on the web. arXiv preprint arXiv:1809.00494, 2018.

[24] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. Communications of the ACM, 51(12):68–74, 2008.

[25] James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. Credibility assessment in the news: do we need to read. In Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining, pages 799–800, 2018.

[26] Souvick Ghosh and Chirag Shah. Towards automatic fake news classification. Proceedings of the Association for Information Science and Technology, 55(1):805–807, 2018.

[27] A. Gînsca, Adrian Popescu, and Mihai Lupu. Credibility in information retrieval. Found. Trends Inf. Retr., 9:355–475, 2015.

[28] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pages 900–903. IEEE, 2017.

[29] Adrian Groza. Detecting fake news for the new coronavirus by reasoning on the covid-19 ontology. arXiv preprint arXiv:2004.12330, 2020.

[30] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In Proceedings of the 2012 SIAM International Conference on Data Mining, pages 153–164. SIAM, 2012.

[31] Kyle Hamilton. Towards an ontology for propaganda detection in news articles. In European Semantic Web Conference, pages 230–241. Springer, 2021.

[32] Amin Ul Haq, Jian Ping Li, Jalaluddin Khan, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad, Ghufran Ahmad Khan, and Amjad Ali. Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors, 20(9):2649, 2020.

[33] Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L Sporer. Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. Personality and social psychology Review, 19(4):307–342, 2015.

[34] Gustav Herdan. A new derivation and interpretation of yule's 'characteristic'k. Zeitschrift für angewandte Mathematik und Physik ZAMP, 6(4):332–339, 1955.

[35] Carla W Hess, Kelley P Ritchie, and Richard G Landry. The type-token ratio and vocabulary performance. Psychological Reports, 55(1):51–57, 1984.

[36] Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In Proceedings of the International AAAI Conference on Web and Social Media, volume 11, 2017.

[37] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2), 2018.

[38] Tao Jiang, Jian Ping Li, Amin Ul Haq, and Abdus Saboor. Fake news detection using deep recurrent neural networks. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pages 205–208. IEEE, 2020.

[39] Tao Jiang, Jian Ping Li, Amin Ul Haq, Abdus Saboor, and Amjad Ali. A novel stacking approach for accurate detection of fake news. IEEE Access, 9:22626–22639, 2021.

[40] Wendell Johnson. Studies in language behavior: A program of research. Psychological Monographs, 56(2):1–15, 1944.

[41] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. Multimedia Tools and Applications, pages 1–24, 2021.

[42] Nir Kshetri and Jeffrey Voas. The economics of "fake news". IT Professional, 19(6):8–12, 2017.

[43] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In Proceedings of the 25th international conference on World Wide Web, pages 591–602, 2016.

[44] Yasmine Lahlou, Sanaa El Fkihi, and Rdouan Faizi. Automatic detection of fake news on online platforms: A survey. In 2019 1st International Conference on Smart Systems and Data Science (ICSSD), pages 1–4. IEEE, 2019.

[45] Yang Liu and Yi-Fang Brook Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In Thirty-second AAAI conference on artificial intelligence, 2018.

[46] Amr Magdy and Nayer Wanas. Web-based statistical fact checking of textual documents. In Proceedings of the 2nd international workshop on Search and mining user-generated contents, pages 103–110, 2010.

[47] Cédric Maigrot. Détection de fausses informations dans les réseaux sociaux. PhD thesis, Université de Rennes 1 [UR1], 2019.

[48] Cédric Maigrot, Ewa Kijak, and Vincent Claveau. Détection de fausses informations dans les réseaux sociaux: l'utilité des fusions de connaissances. 2017.

[49] David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. Lexical diversity and language development. Springer, 2004.

[50] Heinz-Dieter Mass. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. Zeitschrift für Literaturwissenschaft und Linguistik, 2(8):73, 1972.

[51] Svitlana Mazepa, Serhiy Banakh, Andriy Melnyk, Sergiy Pugach, Oleksandra Yavorska, and Natalia Golota. An ontological approach to detecting fake news in online media. In 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), pages 531–535. IEEE, 2021.

[52] Damian Mrowca, Elias Wang, and Atli Kosson. Stance detection for fake news identification. Stanford University, California, US, rep., 2017.

[53] Feyza Altunbey Ozbay and Bilal Alatas. Adaptive salp swarm optimization algorithms with inertia weights for novel fake news detection model in online social media. Multimedia Tools and Applications, 80(26):34333–34357, 2021.

[54] Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. Content based fake news detection using knowledge graphs. In International semantic web conference, pages 669–683. Springer, 2018.

[55] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. arXiv preprint arXiv:1708.07104, 2017.

[56] Stephen R. Pfohl. Stance detection for the fake news challenge with attention and conditional encoding. 2017.

[57] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:1702.05638, 2017.

[58] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 conference on empirical methods in natural language processing, pages 2931–2937, 2017.

[59] Victor Raskin. Linguistics and natural language processing. Machine translation: Theoretical and methodological issues, pages 42–58, 1987.

[60] Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. A simple but tough-to-beat baseline for the fake news challenge stance detection task. arXiv preprint arXiv:1707.03264, 2017.

[61] Arne Roets et al. 'fake news': Incorrect, but hard to correct. the role of cognitive ability on the impact of false information on social impressions. Intelligence, 65:107–110, 2017.

[62] Victoria L Rubin, Yimin Chen, and Nadia K Conroy. Deception detection for news: three types of fakes. Proceedings of the Association for Information Science and Technology, 52(1):1–4, 2015.

[63] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 797–806, 2017.

[64] Danish Shakeel and Nitin Jain. Fake news detection and fact verification using knowledge graphs and machine learning.

[65] Saeid Sheikhi. An effective fake news detection method using woa-xgbtree algorithm and content-based features. Applied Soft Computing, 109:107559, 2021.

[66] Baoxu Shi and Tim Weninger. Fact checking in heterogeneous information networks. In Proceedings of the 25th International Conference Companion on World Wide Web, pages 101–102, 2016.

[67] Prashant Shiralkar, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Finding streams in knowledge graphs to support fact checking. In 2017 IEEE International Conference on Data Mining (ICDM), pages 859–864. IEEE, 2017.

[68] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1):22–36, 2017.

[69] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pages 312–320, 2019.

[70] Michael Siering, Jascha-Alexander Koch, and Amit V Deokar. Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts. Journal of Management Information Systems, 33(2):421–455, 2016.

[71] Edward H Simpson. Measurement of diversity. nature, 163(4148):688–688, 1949.

[72] Kelly Stahl. Fake news detection in social media. California State University Stanislaus, 6:4–15, 2018.

[73] Fiona J Tweedie and R Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. Computers and the Humanities, 32(5):323–352, 1998.

[74] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 18–22, 2014.

[75] Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In Companion Proceedings of the The Web Conference 2018, pages 575–583, 2018.

[76] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017.

[77] Raymond E Wright. Logistic regression. 1995.

[78] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In Proceedings of the eleventh ACM international conference on Web Search and Data Mining, pages 637–645, 2018.

[79] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. Toward computational fact-checking. Proc. VLDB Endow., 7(7):589–600, March 2014.

[80] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 5644–5651, 2019.

[81] C Udny Yule. The statistical study of literary vocabulary. Cambridge University Press, 2014.

[82] Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. Fake news research: Theories, detection strategies, and open problems. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, KDD '19, page 3207–3208, New York, NY, USA, 2019. Association for Computing Machinery.

[83] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. arXiv preprint arXiv:1509.01626, 2015.

[84] Xinyi Zhou and Reza Zafarani. Fake news: A survey of research, detection methods, and opportunities. arXiv preprint arXiv:1812.00315, 2, 2018.

[85] Xinyi Zhou and Reza Zafarani. Network-based fake news detection: A pattern-driven approach. ACM SIGKDD Explorations Newsletter, 21(2):48–60, 2019.

[86] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5):1–40, 2020.

[87] Miodrag Zivkovic, Catalin Stoean, Aleksandar Petrovic, Nebojsa Bacanin, Ivana Strumberger, and Tamara Zivkovic. A novel method for covid-19

pandemic information fake news detection based on the arithmetic optimization algorithm. In 2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), pages 259–266. IEEE, 2021.