International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 61-67

Fake News detection system using TF-IDF

Mrs T. Nathezhtha
Department of CINTEL
SRM Institute of Science and Technology

Dewashish Tarale
Department of Computer Science and Engineering specialization in Software and Engineering
SRM Institute of Science and Technology

Parth Chauhan
Department of Computer Science and Engineering specialization in Software and Engineering
SRM Institute of Science and Technology

Abstract—
Rapid and a vast increase in fake news, defined as probably incorrect information spread with the goal of fraud.Spread of this misinformation is a severe danger that can cause political polarisation. Fake news can make people their political parties or hate a particular community. Thus, fake news is a phenomenon that significantly impacts people's social lives.This system will use NLP techniques to detect fake and misleading news stories from non-reputable sources. We apply the term frequency-inverse document frequency (TF-IDF) of bi-grams and probabilistic context-free grammar (PCFG) detection to a corpus of about 11,000 articles. We find that TF-IDF of bi-grams fed into a Stochastic Gradient Descent model performs exceptionally well at identifying non-credible sources, with PCFGs having slight effects on recall. However, we are sceptical about the generalizability of these findings and include ample discussion on the next steps for exploration in this space.

Date of Submission: 19-04-2023 Date of acceptance: 03-05-2023

I. INTRODUCTION

Around the world almost every body join social media irrespective of age and gender. This helps internet to collect data over vast range of interest of people. People meet new friends sometimes of their interest on social media, and it has a huge potential to spread lots of news and include many people participating through comments, tweets, etc. People have a right to express their opinions and could agree or disagree on specific topics.

Despite the immense potential of online social media, this technology is misused to execute several undesirable acts, such as generating spam, rumors, fake messages, and fake accounts, gaining more substantial influence, creating chaos, or destabilizing homeland security.

It is not like fake news is new, but it got more powerful or rather destructive after the 2016 US election. Due to powerful technology and almost unrestricted access to the internet, news spread faster than earlier. It can reach more people in a concise time. Due to social media, it has become even easy to connect to the whole world and pass and absorb information. Modern Technology had made people believe social media is their primary source of daily news. Research has been done from a journalist's point of view. They could provide analysis and compare the latest news, it asks some six basic questions called 5W+H(what, when, where, who, why, and how)and should present in descending priority from what is most important to the finer details. It would help them to recheck every piece of information before they could publish it. To reduce the spread of questionable content and to make sure that readers are aware of what they are reading, the spread of fake news needs to stop. In the ongoing Pandemic situation, people are more prone to fake news, which could negatively impact society

II. Literature review

Abdullah et al. aimed to develop an optimized convolutional neural network model (OPCNN-FAKE) to detect fake news. The researchers selected this approach because CNNs are effective in natural language processing tasks, including text classification. They believed that optimizing the architecture of the CNN model could improve its performance in detecting fake news.

To evaluate the performance of their optimized model, the researchers compared it with other machine-learning techniques commonly used for fake news detection. These included recurrent neural networks, Long-

term memory, decision tree, logistic regression, K nearest neighbour, random forest, support vector machine, and naïve Bayes. They used four benchmark datasets of fake news to evaluate the performance of each model.

The researchers found that OPCNN-FAKE had the highest performance among all the models based on accuracy, precision, recall, and f1 score. This suggests that the optimized CNN model is effective in detecting fake news.

Accuracy measures how often the model correctly classifies fake and real news articles. Precision measures how often the model correctly typed articles as fake news among all articles it predicted to be fake news. Recall measures how often the model correctly ranked fake news articles among all the fake news articles. The F1 score is a weighted average of precision and recall, often used to evaluate the overall performance of a classification model.

In conclusion, Abdullah et al.'s study prove that optimizing CNN models can lead to more effective fake news detection. Their findings can help inform the development of more accurate and reliable phoney news detection systems in the future.

Jing et al. aimed to develop a method for detecting rumours from microblogs using recurrent neural networks (RNNs). RNNs are a type of neural network that can capture the variations of contextual information of relevant posts over time, making them well-suited for analyzing temporal data such as social media posts.To evaluate the performance of their proposed method, the researchers conducted experiments using 5 million postings collected from Twitter and Sina Web microblogs. They compared the performance of their RNN method with several other machine learning techniques commonly used for rumour detection, including decision trees (DT), random forests (RF), support vector machines (SVM), and two types of RNNs, gated recurrent units (GRUs), and long short-term memory (LSTM). The results of the experiments showed that the RNN method, with LSTM and GRU units, outperformed all other methods and achieved state-of-the-art performance in detecting rumours from microblogs. The researchers concluded that RNNs are effective for detecting words from microblogs, as they can capture the temporal dependencies of the data.

However, to further improve the performance of their proposed method, the researchers suggested adding multiple hidden and embedding layers. Hidden layers are neurons in a neural network that are not directly connected to the inputs or outputs but instead help capture more complex relationships between the inputs and outputs. On the other hand, embedding layers are used to transform the input data into a lower-dimensional space that is easier to analyze. By adding multiple hidden and embedding layers to their RNN method, the researchers believe that they can capture more complex patterns in the data and improve the performance of their method even further. Overall, Jing et al.'s study provide evidence that RNNs can be useful for detecting rumours from microblogs. Further research could lead to even more effective methods for rumour detection in social media.In this paper they 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). After applying Bi-normal separation feature scaling (BSN), a precision of 0.958 is achieved.

Natali et al. proposed a deep hybrid model for detecting fake news called the Capture-Score-Integrate (CSI) model. The CSI model consists of three modules that classify articles as fake or not: the Capture module, the Score module, and the Integrate module. The first module, the Capture module, uses Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) to capture the temporal patterns of user activity on a given article. By analyzing user behaviour patterns, the model can extract mundane ways that indicate whether an article is likely to be fake or not. The second module, the Score module, uses a fully connected neural network layer to capture the characteristics of user behaviour that indicate whether an article is fake or not. This module learns the source characteristics based on the behaviour of users, such as the number of shares or the time spent reading an article. By analyzing these characteristics, the Score module can assign a score to each piece based on its likelihood of being fake. Finally, the Integrate module combines the results from the Capture and Score modules to classify articles as fake. This module takes the scores generated by the Score module and combines them with the mundane patterns extracted by the Capture module to decide whether an article is fake or not. Experimental analysis using real-world data showed that the CSI model achieved higher accuracy than existing models. The model captured meaningful latent representations of users and reports, allowing it to detect fake news articles better. Overall, the CSI model proposed by Natali et al. provides a promising approach to seeing fake news using deep learning techniquesVignesh and his team proposed their algorithm using single shot detector (SSD). Their paper described a method for solving the problem using the SSD model. SSD model is a mainstream algorithm for object detection. It does both image segmentation and image classification simultaneously by running at a single runtime. The SSD method will separate the images. This model will access the bounding box region of the motorcycle and the rider with just one unit. After the area has been picked, the proposed model will classify whether the biker is wearing or not wearing a helmet in real-time. The Convolutional Neural Network is used to recognition.

Kashyap et al. proposed an end-to-end neural network model called "Declare" for debunking fake news and false claims using evidence-aware deep learning. The main idea behind the Declare model is to use evidence gathered from the internet to support or reject a claim. The authors trained a bidirectional Long Short-Term Memory (LSTM) model with at least four different datasets collected from various sources, including social media platforms and fact-checking websites. The datasets contained both true and false claims and evidence to support or reject them.The Declare model works in three stages: claim extraction, evidence retrieval, and verification. In the claim extraction stage, the model uses natural language processing techniques to identify and extract the claim from a text. In the evidence retrieval stage, the model searches for evidence on the internet that either supports or refutes the claim. Finally, in the claim verification stage, the model uses the evidence to determine whether the claim is true or false.The authors evaluated the performance of the Declare model using a test dataset and achieved an overall classification accuracy of 80%. They also compared the performance of their model with other state-of-the-art models for detecting fake news and found that their model outperformed them in terms of accuracy.Overall, the Declare model proposed by Kashyap et al. provides a promising approach to debunking fake news and false claims using evidence-aware deep learning techniques. By incorporating evidence from the internet into the model, the Declare model can make more informed decisions about the veracity of a given claim.

Etal Authors proposed an attention-based Long Short-Term Memory (LSTM) network that uses tweet text, and thirteen different linguistic and user features to distinguish between rumour and non-rumour tweets. The main idea behind the attention-based LSTM network is to use attention mechanisms to selectively focus on certain parts of the tweet text that are more relevant for rumour detection. To evaluate the performance of their model, the authors compared the attention-based LSTM network with various conventional machine learning models (such as Decision Trees, Random Forests, and Support Vector Machines) and other deep learning models (such as Convolutional Neural Networks and Recurrent Neural Networks). The experiments were conducted using a dataset of tweets related to real-world events, labelled as either rumour or non-rumour by expert annotators. The thirteen features used in the model included linguistic features such as sentiment, subjectivity, and readability, as well as useful features such as user reputation and the number of followers. The results showed that the attention-based LSTM network achieved the best performance among all the models evaluated, with an accuracy of over 90%. The attention mechanism was particularly effective in identifying the most informative parts of the tweet text for rumour detection. Overall, the attention-based LSTM network proposed by the authors provides a promising approach for rumour detection on social media platforms. The model can capture linguistic and social cues relevant to rumour detection by incorporating tweet text and user features.

Etal The paper proposed a novel approach for detecting fake news using a TI-CNN model that can handle text and image inputs. The model was designed to capture the correlation between textual content and visual features in detecting fake news. The researchers used two datasets to evaluate the performance of their model and compared it with several other models, including LSTM, CNN, and GRU. The results of the experiments showed that the TI-CNN model outperformed the other models in terms of precision, recall, and F1 measure. This demonstrated the effectiveness of the proposed approach in detecting fake news using both textual and visual information. The TI-CNN model consists of two main components: text and image modules. The text module is a traditional CNN model that takes the textual content of news articles as input and extracts features from them. The image module is also a CNN model that takes the associated images of the news articles as input and extracts visual features from them.The extracted features from text and image modules are combined using a fusion layer, which learns the correlation between the textual and visual elements. The fused parts are then fed into a classification layer, which predicts whether the news article is fake or real. Overall, the TI-CNN model proposed in this paper is a promising approach for detecting fake news that combines textual and visual information. The results of the experiments showed that this model could achieve high performance in detecting fake news, which is important for preventing the spread of misinformation in the digital world.

Etal The paper proposed a novel approach for fake news detection in social networks by focusing on the credibility of the news source. The authors utilized a pattern-driven system, identifying patterns of network behaviour indicative of fake news. Specifically, they analyzed the network of actors involved in the dissemination of information and the credibility of those actors. To evaluate their approach, the authors used two datasets: PolitiFact and BuzzFeed. These datasets contain labelled examples of both fake and real news. The proposed system was compared to several state-of-the-art methods for counterfeit news detection. The results showed that the proposed approach achieved significantly higher accuracy, with an accuracy of 93.30%. Overall, this approach is significant because it addresses the problem of fake news detection from a different perspective by analyzing the network of actors involved in disseminating news. By focusing on the credibility of the sources, this approach can be useful in identifying fake information designed to spread misinformation and propaganda. Early detection of Fake news on social media through propagation path classification with recurrent and convolutional networks, in this paper they used an combination of recurrent and convolutional neural networks to model. News diffusion pathways are multivariate and time series where each tuple of a news story is a numerical vector representing the characteristics of a user 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.

They worked on classification of rumor stance on social media platform with the help of sequential classifiers. In this they use Twitter as their social media platform and describe tweets into 4 categories: 1.Support, 2.Deny, 3.query and 4. comment on an earlier post. They used four sequential classifier-hawkes processes, Long Short Term Memory(LSTM), linear CRF and tree CRF on 8 data sets and all data are related to breaking news.They discover sequential classifiers that use the recitation property in social media interaction outperform non sequential classifiers also LSTM works better than other sequential classifiers.

Here , they worked on rumour detection using NLP and data mining Methods. They defines false news that circulates on social media into two types: long standing rumors and new emerging rumor generate during recent events.They develops a rumor classification system that consist of 4 parts: 1.Detection of rumor, 2.Tracking of rumor, 3. Stances of rumor and 4.Veracity of rumor. And use this system on the PHEME dataset which is publicly available for rumors and non rumors.

They work on increasing the detection of fake news with data imputation. To improve performance they used a novel data preprocessing method to fill the missing value in the raw dataset. With the help of data modeling, they applied missing values for numerical and hierarchical attributes. For hierarchies they select the most frequent value in columns and are numeric for the average value of the column. He did 3 things to cover the missing values. 1. Removed columns with missing values, 2. Missed values with empty text and 3. Used data impersonation techniques to apply missing values and found that multilayer perceptron (MLP) classes improved accuracy by 16%.

TI-CNN is a text and image-based convolutional neural network. They used two datasets to compare their model with several models, such as LSTM, CNN, and GRU. They have demonstrated the effectiveness of this model using real datasets in solving fake news detection problems.The results of these models were very much ahead of other methods in precision, recall and F1 measure.

In this paper, they proposed a network-based pattern-driven approach for fake news detection in social networks. Main idea behind this work is to focus on the credibility of the news source, covering both the bases that create and publish the news and the authorities that spread the word. PolitiFact and BuzzFeed datasets showing were used for good performance compared to the state-of-the-art, with an accuracy of 93.30%

This paper used a combination of recurrent and convolutional neural networks to model.News diffusion pathways are multivariate time series, where each tuple of a news story is a numerical vector representing the characteristics of a user 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%

In 2013, Aditi Gupta, Henmark Lamba and Anupam Joshi achieved more than 90% correct result in identifying false images from twitter of Hurricane Sandy which impacted the United States. They did a characterization analysis to analyze the impact patterns of the fake pictures by analyzing more than 10,000 images on Twitter. They worked on Niave Bayes and Decision tree model. After applying these two ML algorithm they arrives at good result having accuracy of 97% by Decision Tree.

In this they use Twitter as their social media platform and describe tweets into 4 categories: 1.Support, 2.Deny, 3.query and 4. comment on an earlier post. They used four sequential classifier-hawkes processes, Long Short Term Memory(LSTM), linear CRF and tree CRF on 8 data sets and all data are related to breaking news.They discover sequential classifiers that use the recitation property in social media interaction outperform non sequential classifiers also LSTM works better than other sequential classifiers

They work on increasing the detection of fake news with data imputation. To improve performance they used a novel data preprocessing method to fill the missing value in the raw dataset. With the help of data modeling, they applied missing values for numerical and hierarchical attributes. For hierarchies they select the most frequent value in columns and are numeric for the average value of the column. He did 3 things to cover the missing values. 1. Removed columns with missing values, 2. Missed values with empty text and 3. Used data impersonation techniques to apply missing values and found that multilayer perceptron (MLP) classes improved accuracy by 16%.

They purposed the ML algorithm to identify fake news. In this paper three popular methods are used: 1.Naive Bayes 2. Support Vector Machine and 3. Neural Network. They used normalization method for cleaning data so that it works better with correct data. In this paper they found that Naive Bayes has an accuracy of 96.08% and the other two complex techniques has an accuracy of 99.90%.

A. Jain and A. Kasbe work on detecting fake news and they proposed a method so that we can implement this method on Facebook. He used Naive Bayes for forecasting. They used a dataset from Github with 11000 articles divided into (index, text, title and label). Apart from politics, this data contains news related to science and business.For implementation they used both the title and text for their primary source and also added some

references by n-gram And then he compared the results and find that Naïve Bayes (on text with n-grams) gives the accuracy of 0.931 and they also showed some ways to improve this model.

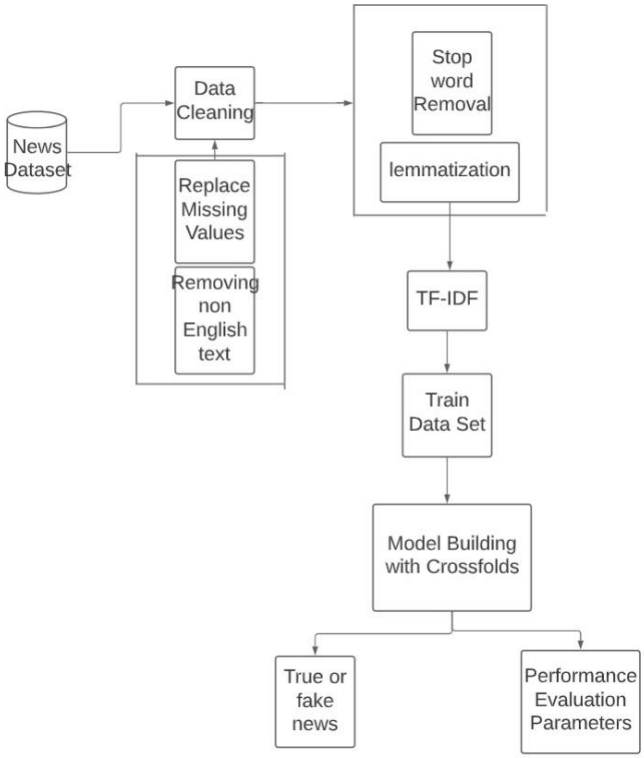
They proposed a model which identified false news tweets from twitter post using combination of (CNN) and (RNN) models.For the dataset they collected 5,800 tweets centered on five rumor stories: Charlie Hebgo, Sydney Siege, Germanwing Crash, Ottawa Shooting, and Ferguson Shooting. Their proposed work on hybrid of CNN & RNN intuitionally identifies important feature related with false news stories without any prior knowledge of news and achieve more than 80% accuracy.

In performance evaluation of fake they divided the dataset of fake news into 2 categories. The first is news and the second is the social context model and they divide news into 2 categories of visual (picture, video) and linguistic (text, title) based.They compared performance between traditional ML methods (Naïve Bayes , Random forest) and the latest deep learning methods (LSTM DROP, LSTM-CNN). The purpose of this paper is to provide a basis so that people can choose between these two approaches. They found that the hybrid CNN - RNN model gives better performance/ results.

They worked on searching on a wide variety of features from news articles, posts and stories that can help predict fake news with greater accuracy. He showed the importance of these new features for the evaluation of fake news. Some of those features are bias, reliability / trustworthiness, engagement, domain location, and temporal patterns. They used a dataset containing 2282 buzzfeed (news articles). They used KNN, Naïve Bayes, Random Forest, Support Vector Machine and XGBoost algorithm for evaluation and to discuss the opportunities and challenges of this approach and they found out that XGBoost work better than all with the accuracy of 0.86.

Here, they classified fake news articles using machine learning models and ensemble techniques (Logistic Regression, Random Forest, Perez- LSVM).In this paper various textual properties are used to differentiate fake new from real news. The experiment was conducted on 4 publicly available dataset which is of different domians and also calculated the performance by performance metrics. The maximum accuracy is 99% achieved by random forest and Perez-LSVM on ISOT Fake News Dataset.

System architecture diagram:

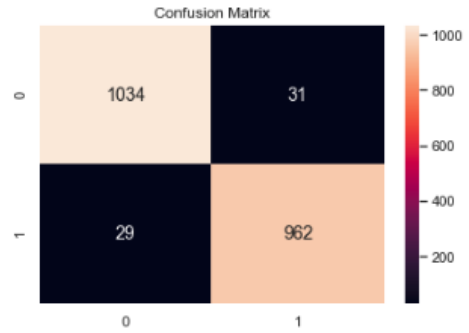


First the data set of the project is taken from the data from excel sheet extracted from the website. Now what will then we will clean and filter the data. A crucial step in building any machine learning model is ensuring the data is properly prepared. This involves cleaning and filtering the data, identifying and removing outliers, and creating independent and meaningful features. By doing so, the resulting model will be more accurate and reliable, as it will be trained on a high- quality dataset free from inconsistencies or irrelevant information. Further details on data preparation techniques and feature engineering will be discussed when working on other machine learning models.

To begin, we will create a TfIdfVectorizer object that will be used to preprocess our text data. This object will be configured to remove common English stop words, which are words that are frequently used in the language but don't provide meaningful information for analysis. Additionally, the maximum document frequency will be set to 0.7, meaning that any terms that occur in more than 70% of the documents will be discarded.

Once we have our vectorizer object set up, we can use it to convert our raw text data into a matrix of Tf-Idf features. Tf-Idf stands for term frequency-inverse document frequency, which is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. By applying this transformation to our data, we can represent each document as a vector of feature values that capture the relative importance of each word in the document. This will allow us to perform more advanced analyses on our text data, such as clustering or classification, and help us uncover insights and patterns that would be difficult to identify otherwise.

III. Results:



Classification Report:	precision	recall	f1-score	support
0	0.97	0.97	0.97	1065
1	0.97	0.97	0.97	991
accuracy			0.97	2056
macro avg	0.97	0.97	0.97	2056
weighted avg	0.97	0.97	0.97	2056

In the process of implementing a machine learning and natural language processing model for identifying fake or factual articles, it is crucial to minimize the number of false positives. This is because false positives can lead to labeling actual facts as fake, which can have significant consequences. In our case, we were able to accurately identify 1034 fake articles and 962 real articles, indicating the effectiveness of our model and the accuracy is 97%.

IV. Conclusion and future enhancement.

This system makes it accessible to everyone. Here are some steps you can consider:

1. Improve the accuracy of the system: The first step to making your system more widely accessible is to make sure that it is accurate and reliable. Consider training the design on a larger dataset, incorporating more advanced machine learning algorithms, or integrating human fact-checking to improve accuracy.
2. Develop a user-friendly interface: Once you have a reliable system in place, you must create a user-friendly interface allowing users to input news articles and receive feedback on their integrity. The interface should be easy to navigate, with clear instructions and feedback on the analysis results.
3. Beta test the system: Before releasing the approach to the public, it's a good idea to beta test it with a small group of users. This will allow you to identify any issues or bugs in the system and make necessary improvements before launching it to a wider audience.
4. Launch the system: Once you have finalized and tested it with a small group of users, it's time to launch it to the public. Consider promoting the system through social media, news outlets, or other channels to reach a wide audience.
5. Collect feedback and make improvements: After launching the system, be sure to collect feedback from users and make any necessary improvements to the system based on their input. This will help you ensure the system remains accurate, reliable, and user-friendly.

References

[1]. A. Gupta, H. Lamba, P. Kumaraguru and A. Joshi, "Faking sandy, identifying fake images of hurricane sandy on twitter," in In proceedings of the 22nd international conference on World Wide Web. ACM, 2013.

[2]. A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik and I. Augenstein, "Discourse-Aware Rumour stance classification in social media using sequential classifier.," Information Processing & Management, vol. 54, pp. 273-290, 2017.

[3]. A. Aker, A. Zubiaga, K. Bontcheva, M. LIakata and R. oricter, "Detection and Resolution of rumours in social media: A survey," ACM comput., vol. 51, p. 36, Feb 2018.

[4]. C. M. M. Kotteti, X. Dong, N. Li and L. Qian, "Fake News Detection Enhancement with Data Imputation," in IEEE 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing, 2018.

[5]. S. Aphiwongsophon and P. Chongstitvatana, "Detecting fake news with Machine Learning methods," in 15th international conference on Electronics, computer, Telecommunication and Information Technology, 2018.

[6]. A. Jain and A. Kasbe, "Fake News Detection," in 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences, 2018.

[7]. O. Ajao, D. Bhowmik and S. Zargari, "Fake news Identification on Twitter with combination of CNN and RNN models," in International conference on social media & society, Copenhagen,Denmark SMSociety, 2018.

[8]. W. Han and V. Mehta, " Han, Wenlin; Mehta, Varshil (2019). [IEEE 2019 IEEE International Conference on Industrial Internet (ICII) - Orlando, FL, USA (2019.1Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation.," in IEEE 2019 IEEE International Conference on Industrial Internet (ICII), Orlando, FL, USA, 2019.

[9]. J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto and E. Cambria, "Supervised Learning for Fake News Detection," in IEEE Intelligent Systems, 2019.