rice2023 Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 57–61
DOI: 10.15439/2023R27
ISSN 2300-5963 ACSIS, Vol. 38

Fake News Identification Using Supervised Machine Learning Algorithms

Md Nooruddin Rabbani
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India
mdnrabbani@gmail.com

Abdul Wahid
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India

Fareeha Rasheed
Department of Computer Science & Information Technology
Maulana Azad National Urdu University
Hyderabad, India

***Abstract*—Fake news has emerged as a significant challenge in today's information-driven society, where misinformation can spread rapidly and have detrimental consequences. Detecting and combatting fake news is crucial in maintaining the integrity of news sources and ensuring the public's access to accurate and reliable information. Machine learning approaches have recently demonstrated the ability to recognize false news stories automatically based on their features and content. To identify fake news, this study compares and contrasts several machine learning (ML) methods, including Random Forest, Passive Aggressive Classifier (PAC), Multinomial Naive Bayes, SVC, Decision tree, Gradient boosting, XG Boost, and Logistic Regression. These algorithms are tested on WELFake_Dataset and the output received has shown a significant increase the accuracy and a decrease in the false rate.**

***Index Terms*—Natural language processing, Passive Aggressive Classifier, Supervised Machine Learning**

I. Introduction

As technology has advanced, social media has become more prevalent in the daily lives of ordinary individuals. People may now regularly consume vast volumes of information from online sources. Fake news has drawn more attention in recent years due to the widespread usage of social media platforms. Popular social media sites like Twitter and Facebook make it simple for users to exchange content, offering them a forum for self-expression and global connectivity. Readers' current initiative is frequently critical to finding solutions to the problem of fake news. Human fact-checking is one of the answers to the issue of false news, which is now a serious concern for both business and academics [2]. Fake news is a fabricated story to deceive readers or spread propaganda. Fake news has become a significant challenge in the age of social media and the internet. Spotting fake news is getting more complicated since more online content is available. The impact on society is that false information may travel swiftly and influence perceptions. It might destroy trust between various social groups, affecting dialogue and decreasing confidence in the media. Due to how easily inaccurate information may be distorted to support a false narrative, it can also result in social unrest and chaos.

During the global COVID-19 outbreak, several doctored films and images about the COVID-19 virus, its origin and spread of vaccines, and the deaths it has caused have been circulating on social media. The percentage of fake news, videos, and photographs disseminated on social media is thought to be between 30 and 35 per cent. This erroneous information causes widespread panic since it travels faster than the virus [17]. As a result, there is an increasing demand for automated fake news detection systems that can tell legitimate news pieces from false ones. To detect more fake news, a machine-learning system is being deployed. Therefore, various ML techniques have been used in this research to identify fake news. Figure 1 shows the ML application to classify fake and real news. The WELFake_Dataset dataset was obtained from Kaggle.

II. Motivation

Fake news is a serious issue that has the potential to compromise democracy and public trust. In recent years, the growth of false information on the internet via social media and websites has caused harm to society by leading people to make incorrect decisions and propagating false information. Like other techniques, machine learning methods are one way to spot fake news. These days, more and more people are interested in using machine learning techniques to detect fake news. During Pendamic, lots of fake news was spreading that motivated me to work on it.

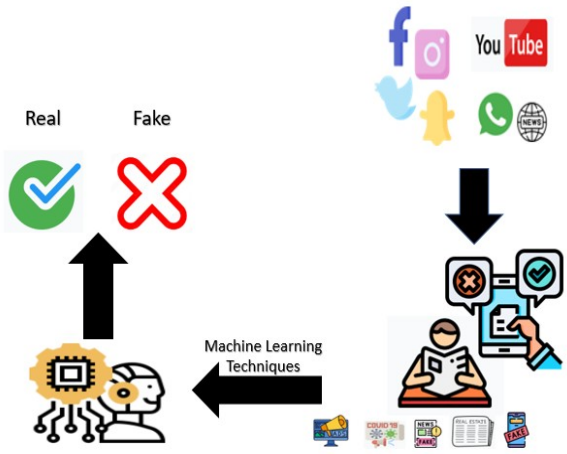

Figure-1 ML application to classify the fake and real news.

III. Related Work

The study by Narra et al. (2022)[1]. presents a selected feature set-based method for identifying COVID-19-related

fake news. The authors chose the essential attributes using approaches like chi-square, mutual information, and information gain. They also compared the performance of different classification algorithms and found that SVM outperformed other algorithms. The study highlights the importance of contextual information in fake news detection and shows that sentiment analysis, readability, and topic modelling are the most compelling features for detecting fake news related to COVID-19. However, the study has limitations, such as a small dataset focusing only on English-language news articles. The study offers insightful information on identifying false news and encourages future study.A fresh approach to identifying false news is put out in the study [2]. Five machine learning models and three deep learning models comprise the stacking ensemble used in the procedure. Two datasets of fictitious and legitimate news stories train the models. Afterwards, the news stories are categorized using the stacking ensemble. In the ISOT and KDnugget datasets, the approach obtains an accuracy of 99.94% and 96.05%. The work [3] proposes a hybrid language and knowledge-based technique for identifying fake news on social media. The strategy combines linguistic characteristics, such as word count, readability, and lexical diversity, with knowledge-based characteristics, such as the website's standing where the news is published, the number of sources used to assemble the news, and fact-checking by reputable fact-checking websites. The method gets a 94.4% accuracy rate when tested on a dataset of actual and fake news articles. The paper [4] suggests a taxonomy of strategies for categorizing fake news. It examines the most recent methods for classifying fake news. and goes over each technique's merits and drawbacks. The three categories that make up the taxonomy are feature type, classification algorithm, and evaluation measure. The paper [5] addresses the issue of identifying false information regarding COVID-19. The World Health Organization, UNICEF, and the United Nations are used as information sources and epidemiological data gathered from various fact-checking websites in the authors' proposed approach for detecting deceptive information. A dataset of 10,000 news stories about COVID-19 is used to assess the model. The findings reveal that the model has a 90% accuracy rate for detecting false information. This paper [6] proposes using the OPCNN-fake, an optimized convolutional neural network, to identify fake news. OPCNN-FAKE is a deep learning model that distinguishes between genuine and fraudulent text via recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for feature extraction. The authors compare the performance of OPCNN-FAKE with other machine learning (ML) and deep learning (DL) models on four fake news benchmark datasets. The findings show that OPCNN-FAKE outperforms the other models on all datasets, with an average accuracy of 92.8%. The authors conclude that OPCNN-FAKE is a promising model. for fake news detection and can be used to improve the accuracy of fake news detection systems. The paper [7] offers a fake media detection system based on blockchain and natural language processing (NLP). The system uses natural language processing (NLP) to extract components from media material, such as using particular words or phrases, logical fallacies, and the overall tone. Then, a machine learning model is trained using these traits to determine if a piece of media is authentic or fraudulent. The findings of the machine learning model are also stored within the system using blockchain technology, making it challenging for attackers to alter or remove the results. The paper's authors evaluate the system on a dataset of fake and authentic media content. The findings reveal that the technology has a 92% accuracy rate for identifying fake material. The paper [8] discusses the problem of fake news detection in social media. Fake news is defined as news that is intentionally false or misleading. It is often created to deceive people and to manipulate public opinion. Fake news can harm society, leading to people making decisions based on false information. The paper discusses several different approaches to fake news detection: content-based, social media-based, and Hybrid approaches. The paper [9] explains how to spot fake news using machine learning algorithms. The feature extraction process entails taking specific information from the news articles, such as the author, text, and title. A machine learning algorithm determines whether the news stories are authentic in the classification stage. On a dataset of 1,000 news articles, the authors assessed their methodology. They discovered that their method had a 95% accuracy rate. The paper [10] offers a hybrid deep-learning architecture for detecting fake news stances. The architecture comprises the long short-term memory (LSTM) network and convolutional neural network. The LSTM is used to record the sequential associations between the features, while the CNN is used to extract features from the news item. A dataset of news items with four stances—agree, disagree, discuss, and unrelated—was used to train the architecture. The experimental findings indicate an accuracy of 97.8%.The paper [11], Using a machine learning tool, the suggested method extracts various texts from the articles and feeds the feature set into the models. The training models were trained, and their parameters were adjusted for the best outcome.

The paper [12] proposes using the OPCNN-fake, an optimized convolutional neural network, to identify fake news. OPCNN-FAKE is a deep learning model that distinguishes between genuine and fraudulent text via recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for feature extraction. The attention module learns to focus on critical features, the feature extraction module extracts feature from the text, and the classification module determines if the news is true or false. A dataset of news items and their labels is used to train the MVAN model. The results of the FNC-1 and FNC-2 datasets demonstrate that the MVAN model performs better than cutting-edge techniques.

IV. Proposed Model

In this is section proposed methodology discussed.

This comparative study comprised various machine learning (ML)classifiers for the dataset shown in this figure-2, collected from the online source Kaggle name-WELFake_Dataset. The planned work's phases are shown in Figure 2 are simplified as follows: Data processing has been done before feature extraction and splits into a ratio of 80:20 for training and test data. ML classifiers such as Random

TABLE-1 RELATED WORK DATASET AND MODEL USED

Ref	Year	Contribution	Dataset	Model used
[12]	2021	Multi-View Attention Networks(MVAN) for Fake News Detection on social media	Twitter15 and Twitter16	Multi-view attention networks (MVAN)
[13]	2023	Fake news detection in social media based on sentiment analysis using classifier techniques.	ISOT false news and LIAR	Naïve Bayes,Passive Aggressive clssifer, DNN
[14]	2021	"All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustness of Fake-news Detectors Under Black-Box Settings".	Kaggle fake-news dataset, ISOT dataset, and LIAR dataset.	MLP, CNN, RNN and Hybrid CNN-RNN
[15[	2020	"Detecting Misleading Information on COVID-19,"	Primary data is collected from various online sources using the Google Fact Check Tools API and stored in MySQL Server.	Decision Tree (DT), KNN Logistic Regression (LR), Linear Support Vector Machines (LSVM), Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Perceptron, Neural Network (NN), Ensemble Random Forest (ERF), Extreme Gradient Boosting classifiers (XGBoost)
[16]	2020	"Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM),"	Fake News Challenges	CNN and LSTM
[17]	2022	Detecting covid-19-related fake news using feature extraction	The primary dataset used was collected from Facebook, Twitter, The New York Times	Random forest classifier (RFC), AdaBoost, DT and KNN

Forest, Passive aggressive classifier, multinomial Naive Bayes, SVC, Decision tree, Gradient boosting, XG Boost, and logistic regression are used to obtain the best accuracy and result.

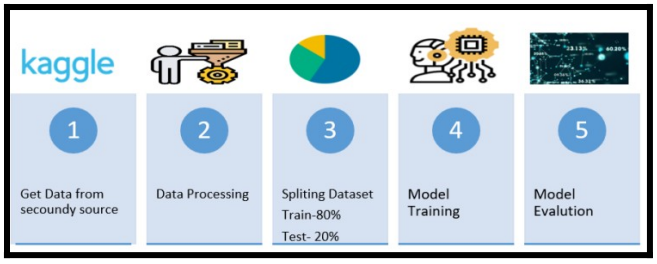

Figure-2 Steps of proposed ML model.

A. Preprocessing

Data processing methods are essential for obtaining insightful information in the age of big data. These methods include removing null values, omitting unnecessary characters, stemming, and vectorization, which helps to improve the quality and usefulness of their datasets, leading to more precise and trustworthy analysis

B. Null Values

Null values, commonly called missing values, can significantly affect data analysis. They may add bias, impact statistical computations, or hinder machine learning techniques. Data preparation requires recognizing and effectively treating null values to solve this problem.

C. Removing Extra Character

In real-world datasets, it is common to encounter unstructured or noisy data containing extra characters, such as punctuation marks, special symbols, or HTML tags. These extraneous elements can interfere with analysis tasks, like text mining or natural language processing.

D. Stemming

Using the text normalization approach known as stemming, data researchers may execute tasks like sentiment analysis, information retrieval, and document clustering more effectively.

E. Vectorization

Vectorization is a fundamental step in transforming textual or categorical data into numerical representations. The process involves converting words, phrases, or categorical labels into numerical vectors that capture the inherent relationships between them. Many machine learning algorithms require numerical input, making vectorization essential.

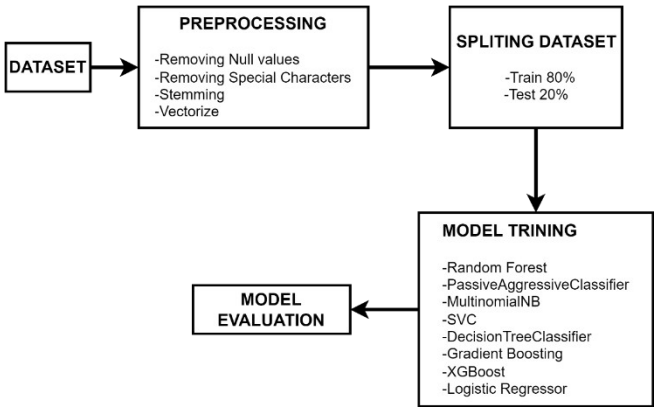

Figure-3 Proposed methodology.

This section provides a detailed explanation of the proposed work. The planned work's steps are depicted in Figure-3 and are summed up as follows:

V. PARAMETERS OF EVALUATION

The effectiveness of categorization models is frequently evaluated using the following metrics [7]

Accuracy- Number of correct predictions to the total number of predictions.

Precision -The precision metric is used to overcome the limitation of Accuracy. The precision determines the propor-

TABLE-2 FORMULA USED FOR PROPOSED METHODOLOGY.

Metric	Formula	Interpretation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Performance of models based on train data
Precision	$\frac{TP}{TP+FP}$	The quality of a positive prediction made by the model
Recall	$\frac{TP}{TP+FN}$	The true positive rate (TPR) model correctly identifies.
F1 Score	$\frac{2\,x\,Precision\,x\,Recall}{Precision+Recall}$	It measures the model's accuracy.
Specificity	$\frac{TN}{FP+TN}$	Measures the proportion of true negatives that the model correctly identifies.

tion of positive prediction that was actually correct. It can be calculated as the True Positive and True Negative.

Recall- It can be calculated as True Positive or predictions that are actually true to the total number of positives.

F-Scores- F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them

Specificity-Specificity measures the proportion of true negatives that are correctly identified by the model.

This experiment on WELFake_Dataset in Jupiter required a system with at least a Core i5 CPU, 16GiB module 120GB SSD free space. As shown in Table-2, the Random Forest algorithm is also a powerful tool for detecting fake news. Compared to other algorithms, it requires less training time, and with careful data preprocessing and model tuning, levels % of this experiment's accuracy of 97% was achieved. Passive-Aggressive algorithms are generally used for large-scale learning, and usually, for large-scale learning, passive-aggressive algorithms are utilized. Nowadays, this is most popular to detect fake news on social media like Twitter, where new data is added every second.

A. Passive-Aggressive Algorithms

Machine learning algorithms, passive-aggressive algorithms (PAA), are frequently employed for binary classification problems. They are renowned for their simplicity and effectiveness, especially in online learning settings where immediate forecasts are necessary, and information is delivered in batches. Nowadays, social media websites like Twitter, where new information is uploaded every minute, are the most widely used for spotting fake news. In this experiment, it achieved a level of accuracy of 95%. Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because

B. Multinomial NB

Naive Bayes is a classification algorithm based on Bayes' theorem, assuming independence among the features. It is widely used for text classification, spam filtering, sentiment analysis, and other tasks. Using Bayes' theorem, the approach determines the probability of each class provided the observable characteristics. Similar to the PPA algorithm, an accuracy level of accuracy 95% was achieved.

P (class | features) = (P(class) x P (features | class)) / P(features) [7]

C. SVC

Support Vector Machine with Citation, or SVC, is a powerful supervised learning technique for classification and regression applications. SVCs have been used in many fields, including bioinformatics, text classification, and picture recognition.

The foundation of SVMs is the discovery of a hyperplane or group of hyperplanes in a very high-dimensional or infinite-dimensional space that may be applied to classification, regression, or other tasks like outlier detection. The closest points from each class to the line the algorithm draws to divide the data into classes are referred to as support vectors [19].

The margin is the distance between the line and the support vectors, and maximizing the margin is the objective [20]. Many people like SVMs because they generate substantial accuracy while requiring minimal processing resources.

Accuracy, precision, recall, and F1 score are some of the measures used to assess an SVM's effectiveness in identifying fake news. These metrics offer a quantitative evaluation of the model's capacity to distinguish between instances of fake and legitimate news. SVM has high accuracy and reliable results in this research work, and the performance of SVM is high as compared to other ML algorithms used that achieve a high level of accuracy of 97% on this dataset.

D. Decision Tree

The decision tree (DT) classifier is one of the most popular ML algorithms for classification and prediction problems on supervised data. It offers a 96% accuracy rate for the dataset used. Using rules and trees, the training dataset is segmented into classes.

E. Gradient Boosting

Gradient boosting is another ensemble method used in machine learning to strengthen weak learners. 96% accuracy level on this dataset was attained.

F. XGBoost

Extreme Gradient Boosting, sometimes called XGBoost, is a popular machine learning method that performs very well in various structured data applications, including classification, regression, and ranking problems. On this dataset, it has a 96% accuracy rate.

G. Logistic Regression

Logistic regression is a prominent and widely used classification method for categorizing fake news. Logistic regression is a statistical technique for binary classification tasks. It predicts the chance that an instance will belong to a specific class based on the values of the input characteristics. It also has a 97% accuracy rate for this dataset.

TABLE-2 ACCURACY OF ML ALGORITHMS.

Algorithm	Predication Accuracy	Precision	Recall	F1 Score
Random Forest	96	100	95.917023	97.93638
Passive Aggressive Classifier	95	100	95.164979	97.92024
MultinomialNB	95	100	95.164979	97.92024
SVC	97	100	96.841755	97.95588
Decision Tree	96	99	96.419098	96
Gradient boosting	96	99	96.692282	96
XGBoost	96	72	33	45
Logistic Regression	97	91	43	59

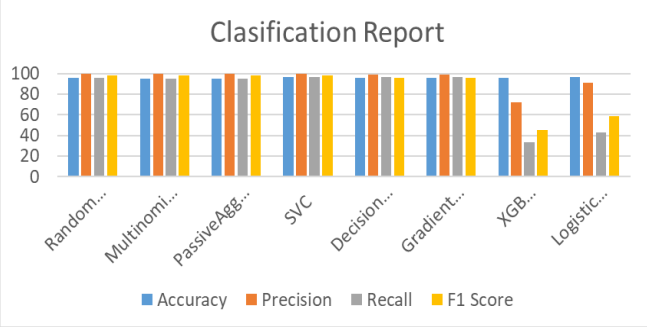

Figure-4 Classification Report

VI. CONCLUSION

Machine learning techniques offer a promising solution for detecting and preventing the spread of fake news. Our model approach to identifying fake news using supervised ML techniques uses natural language processing techniques to analyze the text of news articles and identify patterns and features associated with fake news. This study shows that SVM is a highly effective algorithm for spotting false information. It is a valuable tool in the fight against the spread of false information and disinformation in the digital age because of its incredible accuracy, precision, recall, and low false rate. The findings of this study can be used as a basis for the creation of sophisticated and scalable false news detection systems that can help uphold the reliability and integrity of online information. Future research has several opportunities to enhance further the functionality and application of this method in this field.

REFERENCES

[1] Narra, M., Umer, M., Sadiq, S., Eshmawi, A. A., Karamti, H., Mohamed, A., & Ashraf, I. (2022). Selective feature sets based fake news detection for covid-19 to manage Infodemic. IEEE Access, 10, 98724–98736. https://doi.org/10.1109/access.2022.3206963

[2] T. Smith, P. Husbands and M. O'Shea, "Neutral networks in an evolutionary robotics search space," Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), Seoul, Korea (South), 2001, pp. 136-143 vol. 1, doi: 10.1109/CEC.2001.934382.

[3] Seddari, Noureddine & Derhab, Abdelouahid & Belaoued, Mohamed & Halboob, Waleed & Al-Muhtadi, Jalal & Bouras, Abdelghani. (2022). A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media. IEEE Access. 1-1. 10.1109/ACCESS.2022.3181184.

[4] Rohera, Dhiren & Shethna, Harshal & Patel, Keyur & Thakker, Urvish & Tanwar, Sudeep & Gupta, Rajesh & Hong, Wei-Chiang & Sharma, Ravi. (2022). A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects. IEEE Access. 10. 10.1109/ACCESS.2022.3159651.

[5] M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[6] H. Saleh, A. Alharbi and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," in IEEE Access, vol. 9, pp. 129471-129489, 2021, doi: 10.1109/ACCESS.2021.3112806.

[7] Shahbazi, Zeinab & Byun, Yungcheol. (2021). Fake Media Detection Based on Natural Language Processing and Blockchain Approaches. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3112607.

[8] Rohera, Dhiren & Shethna, Harshal & Patel, Keyur & Thakker, Urvish & Tanwar, Sudeep & Gupta, Rajesh & Hong, Wei-Chiang & Sharma, Ravi. (2022). A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects. IEEE Access. 10. 10.1109/ACCESS.2022.3159651.

[9] Stahl, K. (2018, April 20). Fake news detection in social media. California State University Stanislaus. Retrieved May 15, 2018.

[10] Sahu, Mickey, and Narendra Sharma. "Research on the Ability to Detect Fake News with Machine Learning." International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 4, 2020, pp. 3664-3668., doi:10.18531/ijert.2020.9.4.1590.

[11] Xu, Y., Zhang, Y., Li, J., & Zhang, J. (2021). MVAN: Multi-view attention networks for fake news detection on social media. IEEE Access, 9, 106907-106917. doi:10.1109/ACCESS.2021.3100245.

[12] Balshetwar, S.V., RS, A. & R, DJ Fake news detection in social media based on sentiment analysis using classifier techniques. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-14883-3

[13] Ali, H., Rehman, S. U., Shah, M. A., & Imran, M. (2021). All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings. IEEE Access, 9, 81678-81692. doi:10.1109/ACCESS.2021.3085875

[14] M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[15] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi and B. -W. On, "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)," in IEEE Access, vol. 8, pp. 156695-156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[16] Khan, S., Hakak, S., Deepa, N., Prabadevi, B., Dev, K., & Trelova, S. (2022). Detecting covid-19-related fake news using feature extraction. Frontiers in Public Health, 9. https://doi.org/10.3389/fpubh.2021.788074

[17] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification (2nd ed.). Wiley-Interscience.

[18] https://en.wikipedia.org/wiki/Support_vector_machine

[19] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[20] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). Fake News Detection on Social Media: A Data Mining Perspective. ACM Computing Surveys, 52(5), 1-34