International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-3, May 2021
https://doi.org/10.21276/ijircst.2021.9.3.8
Article ID IRP1150, Pages 48-52
www.ijircst.org

Media Manipulation Detection System Using Passive Aggressive

Ms. Aarti Chugh, Dr. Yojna Arora, Mr. Jaivardhan Singh, Mr. Shobhit, Mr. Ronak

Department of Computer Science & Engineering, Amity University, Gurugram, Haryana, India

Correspondence should be addressed to Ms. Aarti chugh**;** achugh@ggn.amity.edu

ABSTRACT-Due to the extreme growing use of social media and online news media, there has been a rise in fake news recently. It has become much easier to spread fake news than it was before. This type of fake news, if widely circulated, could have a significant impact. As a result, it is necessary to take steps to reduce or distinguish between true and false news. We design a system to verifying such type of news and extract correct news or provide correct news corresponding to the fake news. On a text-based dataset, we give an overview of false news detection using various classifiers such as Passive Aggressive Classifier, Random forest, Logistic regression and decision tree classifier gets better results, as seen by the work done. Also top ten recommendations corresponding to the real news is displayed through our proposed model.

KEYWORDS- Fake news, machine learning, classification, Authenticity.

I. INTRODUCTION

Fake news can be divided into three categories in general. The first category includes false news, which is news that is entirely made up by the article's writers. The second type of fake news is fake satirical news, which is fake news with the primary goal of amusing readers. The third group consists of poorly published news articles that contain some actual news but are not completely factual. In a nutshell, it's reporting that uses, say, quotes from politicians to report a completely false storey. Often, this type of news is intended to support a certain agenda or a skewed viewpoint**.** As the circulation of misinformation online increases, particularly in media sources such as social media feeds, news websites, and online newspapers, fake news detection has prompted the attention of the general public and researchers. According to a research study by Frederick Burr Opper, Face book referrals accounted for 50% of all traffic to fake news pages and 20% of all traffic to major websites. Since 62 percent of American adults receive their news from social media, being able to detect fake stories in web media is critical. Fake profiles, fake tweets, and false news threaten social media and the internet. The aim is frequently to mislead readers and/or trick them into buying or accepting something that isn't true. As a result, a machine like this could help solve a problem to some extent.

II. LITERATURE REVIEW

In their paper, Mykhailo Granik et al. [3] demonstrate a basic method for detecting false news using a naive Bayes classifier. This method was turned into a software framework and put to the test on a series of Facebook news messages. They came from three major Facebook groups on the right and left, as well as three major national political news sites (Politico, CNN, and ABC News). They were able to obtain a classification accuracy of about 74%. The accuracy of false news classification is marginally lower. This may be due to the dataset's skewness: Just 4.9 percent of it is false information. Himank Gupta et al. [4] proposed a framework based on various machine learning initiatives to understanding a variety of issues, including lack of precision, time lag (Bot Maker), and fast processing time to manage thousands of tweets in one second. To begin, they gathered 400,000 tweets from the HSpam14 dataset. The 150,000 spam tweets and 250,000 non-spam tweets are then further listed. They also derived some lightweight features from the Bag-of-Words model, and also the Top-30 words that have the most detail gain. 4. They were able to achieve an accuracy of 91.65%, surpassing the previous solution by nearly 18%. Marco L. Della Vedova et al. [5] proposed the first machine learning (ML) false news identification system, which outperforms current approaches in the literature by integrating news information and social background functionality, raising accuracy to 78.8%. Second, they applied their approach in a Face book Messenger Chabot and tested it with a real-world sample, achieving an 81.35 percent accuracy in detecting false news. Their aim was to evaluate whether a news storey was accurate or not; they first explained the datasets they used, then introduced the content-based approach they used and the tool they proposed to integrate it with an existing social-based approach. The resulting data consists 15,500 articles from 32 different pages (14 conspiracy pages, 18 scientific pages) 900,000+ people have sent it over 2,300,000 likes. Hoaxes account for 8,923 (57.6%) of the messages, while non-hoaxes account for 6,577 (42.4%). Cody Buntain et al. [11] learn to forecast accuracy tests in two credibility-focused Twitter datasets: CREDBANK, a crowd-sourced dataset of accuracy evaluations for incidents on Twitter, and PHEME, a dataset of possible rumours on Twitter and journalistic estimates of their accuracies. They use Twitter material from Buzz Feed's fake news dataset to implement this tool. A feature analysis describes the characteristics that are most predictive for crowd-sourced and journalistic accuracy tests, with the findings correlating with previous research. They classify stories by defining heavily retweeted message threads and using the characteristics of these threads to classify stories, restricting the applicability of this work to a subset of common tweets. This approach

can only be used on a small percentage of Twitter message threads since the rest of messages are barely retweeted.

A. Proposed Work and Methodology

The method, which is built in three sections, is explained in this article. The first section is static and is based on a machine learning classifier. We investigated and conditioned the model with four different classifiers before deciding on the right one for final implementation. The second component is complex, and it takes the user's keyword/text and scans the internet for the news' truth chance.

We used Python and its Sci-kit libraries in this paper. Python offers a large library of extensions and libraries that can be used in Machine Learning. The Sci-Kit Learn library is the best source for machine learning algorithms, as it provides access to nearly all types of machine learning algorithms for Python, making for fast and rapid evaluation of ML algorithms. We used a GUI for the model's prediction-based deployment, which allows for client-side execution by pasting the text. We've also used API requests to scrape data from the internet.

B. System Design

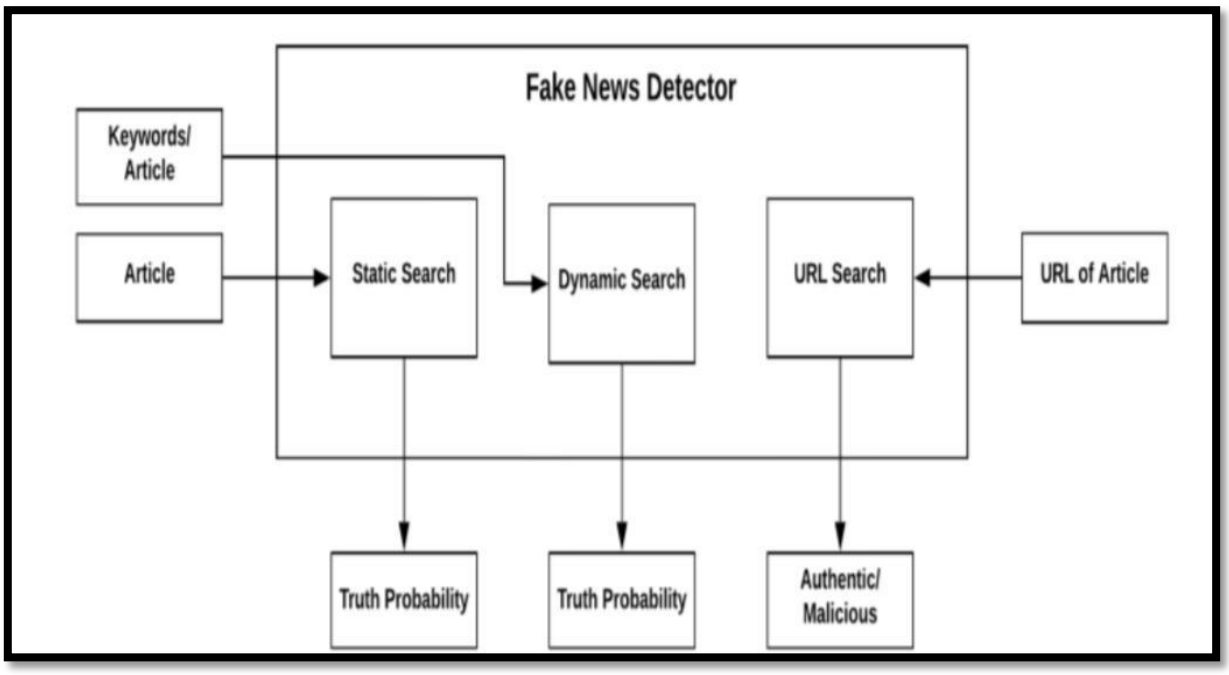

Fig. 1: System design source fake news detector

C. System Architecture

The static part of the fake news detection systems architecture is clear and follows the basic machine learning algorithm flow. The machine architecture is self-explanatory and can be seen in figure 2 shown below.The below are the core production processes:

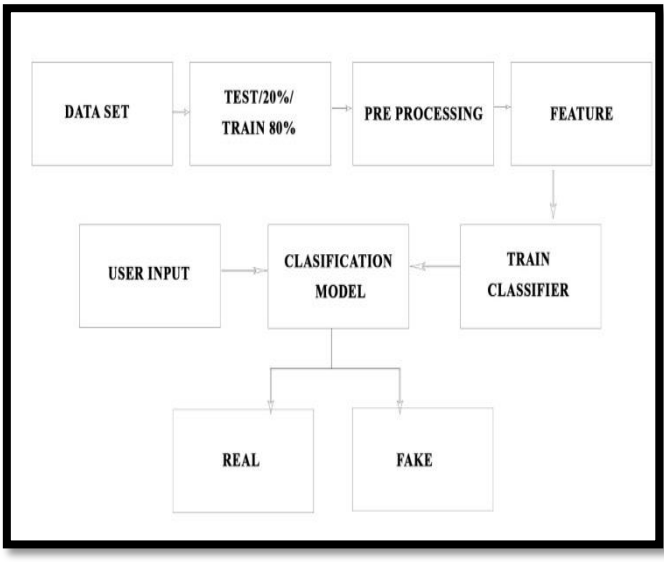

Fig. 2: Proposed System Architecture

D. Dynamic Search

The site's second search field requests unique keywords to be searched on the database, after which it generates an appropriate output for the true and false probability of that term being found in an article or a related article with such keyword references in it.

E. GUI Search

The site's third search area recognises a single website domain name, after which the implementation searches our true sites database or our blacklisted content database for the article. The true article directory contains domain names that offer proper and authentic news on daily basis, and vice versa. If the domain isn't found in any database, the implementation doesn't classify it; rather, it simply states that the news aggregator doesn't exist.

F. Logistic Regression

Logistic regression is a form of statistical analysis that uses prior observations to predict the outcome of a dependent variable. Its primary use is binary classification. A logistic regression algorithm examines the relationship between one or more dependent variables and a dependent variable. For higher accuracy, logistic regression requires a large data set, while Naive Bayes will operate with a smaller data set

G. Decision Tree

By transforming data into a tree representation, Decision Tree solves the issue of machine learning. Every attribute is represented by an internal node in the tree representation, and every class mark is represented by a leaf node. Both regression and classification problems are often solved using decision tree algorithms. When there are a lot of sparse elements, a decision tree can over fit.

H. Random Forest

Often referred to as Random Decision Forests, Random Forests may be used for classification and regression issues. This also can be used in the unsupervised technique. The Random Forest technique was introduced by Brieman. Predictions of several trees are combined by random forest classifiers. Many decision trees are built by the random forest algorithm. Utilizing a subset of features, each decision tree is created. Each decision tree produces one class and eventually bootstraps the votes to obtain the better accuracy from the Random Forest technique.

I. Passive Aggressive Classifier

The passive-aggressive algorithms are a class of large-scale learning algorithms. They, like the Perception, do not need a learning rate. They do, however, have a regularization parameter, unlike the Perception

Table 1: Passive Aggressive Classifier

Algorithm	Random Forest	Logistic Regression	Decision Tree	Passive Aggressive Classifier
Accuracy	99.01%	98.81%	99.69%	92.50%

III. IMPLEMENTATION

A. Static Search Implementation

In static part, we have trained and used 2 out of 4 algorithms for classification. They are Random Forest and Logistic Regression.

Step 1: We extracted features from the pre-processed dataset in the first stage. Bag-of-words, Tf-Idf Functions, and N-grams are examples of these features.

Step 2: We've created all of the classifiers for predicting the detection of fake news here. Different classifiers are fed the extracted features. We used sklearn's Logistic Regression and Random Forest Classifiers. All of the classifiers used each of the extracted features.

Step 3: We compared the f1 score and tested the confusion matrix after fitting the model.

Step 4: The two best-performing models were chosen as candidate models for detecting fake news after fitting all of the classifiers.

Step 5: We used Grid Search CV methods to perform parameter tuning on these candidate models and choose the best performing parameters for these classifiers.

Step 6: Finally, the chosen model was used to identify false news using the likelihood of fact.

Step 7: Logistic Regression was the final and highest performing classifier, and it was saved to disc. It can be used to categorise and identify fake news.

It takes a user-supplied news article as input, then applies the model to produce the final classification output, which is shown alongside the likelihood of truth to the user.

B. Dynamic Search Implementation

Dynamic implementation contains 3 search fields which are-

1. Search by article content.
2. Search using key terms.
3. Search for website in database

To come up with a proper solution for the problem, we used Natural Language Processing in the first search area, and as a result, we attempted to construct a model that can identify fake news based on the words used in newspaper articles. Our programme employs NLP techniques such as Count Vectorization and TF-IDF Vectorization before passing it through a Passive Aggressive Classifier to determine the validity of an article in terms of a percentage probability. The site's second search field requests unique keywords to be searched on the internet, and it returns a suitable percentage likelihood of that term being present in an article or a related article containing such keyword references. The site's third search field recognises a particular website domain name, after which the implementation searches our true sites database or our blacklisted sites database for the site. The true sites database contains domain names that provide proper and authentic news on a regular basis, and vice versa. If the domain isn't contained in either database, the implementation doesn't identify it; instead, it simply states that the news aggregator doesn't exist.

C. Working

The input can be broken down into 3 statements-

To verify the accuracy of a news article, use natural language processing (NLP). If a user has a question about the validity of a search query, we can conduct a direct search on our platform and generate a confidence score using our custom algorithm. Verify the credibility of a news outlet. In our implementation of the problem statement, these parts were created as search fields to take inputs in three different ways. In This GUI you can see that there is a Login Page (Fig 3:). To use this GUI you need to follow 4 Simple Steps. First you have to create an account on this Fake news tool, after that you have to Enter your Credentials. Once you enter your username and Password you will enter inside the GUI. Once you are inside the GUI you can Select algorithm of your choice (Can be seen in Fig 4). Out of all the algorithm's Passive aggressive is the most popular one. Now after choosing the algorithm you have to enter the news link in the "Query text' field. Query Field is the section where you have to enter link of the news you want to check if it's real or not (fig 5). Once you have entered the link just tap on "Check Authenticity). Now the tool will access top 10 news websites automatically and it will match your news link with them. If your query is fake then you Will see " The News is fake " Notification and if its real you will get " The news is authentic' as shown in fig 6

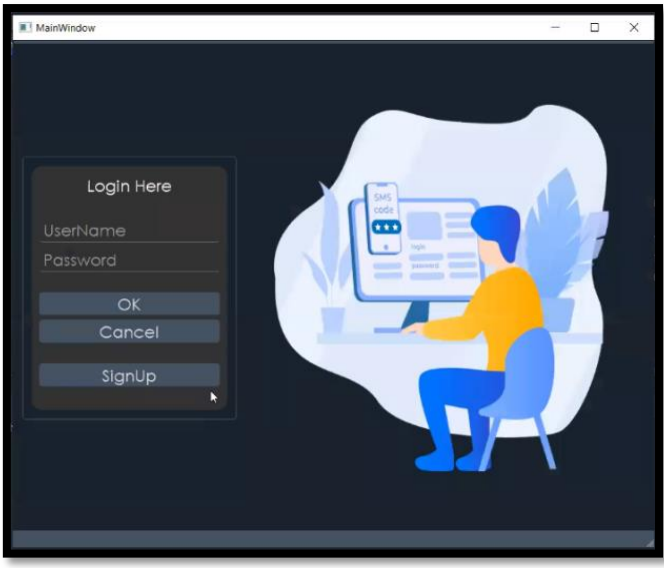

Fig. 3: Login page

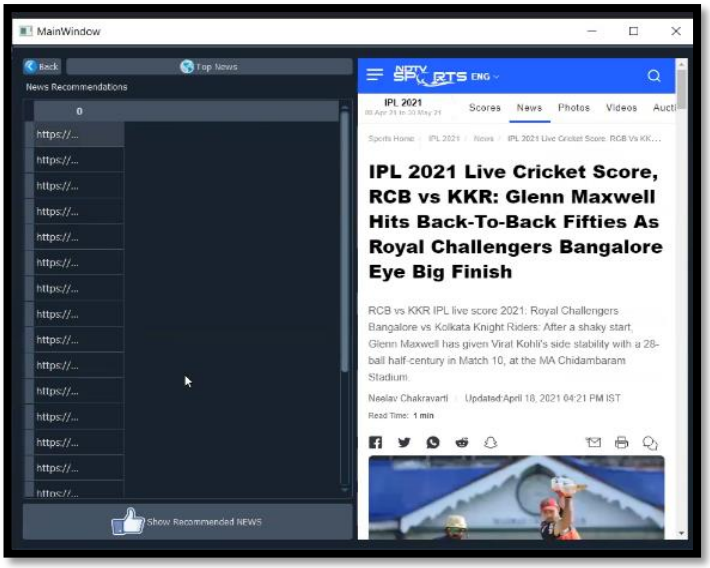

Fig. 6: Shows top ten news

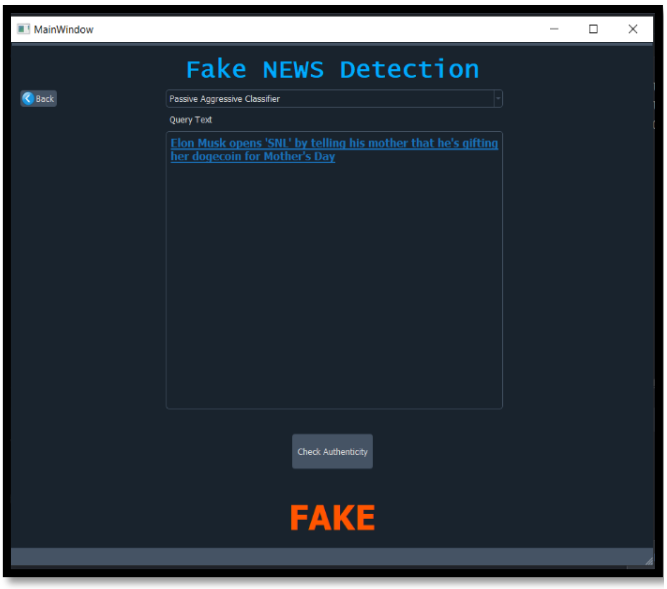

Fig. 4: Pre clearing of News

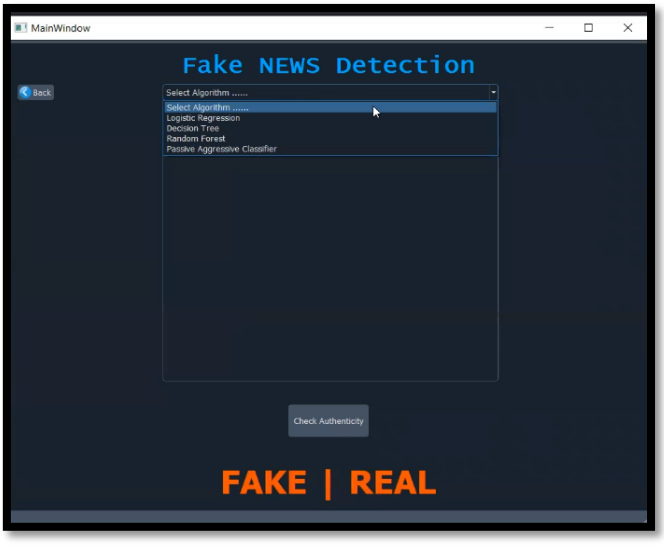

Fig. 5: Numbers of Algorithms

IV. CONCLUSION

The proposed work shows working of four machine learning algorithms namely Logistic Regression, Decision Tree, Random Forest, Passive Aggressive Classifierfor the Kaggle dataset to predict the fake news on social media. The accuracy of prediction using the Logistic Regression algorithm was found to be 98.81%, the Random forest algorithm is capable to predict with an accuracy of 99.01%, Decision Tree is capable to predict with an accuracy of 99.69% and Passive Aggressive Classifier gives the accuracy of 92.50%. After analysis, it is found that the Decision Tree algorithm performs better and can be used efficiently for the detection of fake news. The work can be further extended to huge datasets from other websites that contain a greater number of social media websites and networks. Other different algorithms can also be used in combination to achieve a greater accuracy of prediction.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema –"Fake News Detection on Machine Learning and Natural LanguageProcess".
[2] Kai Shu , Amy Sliva , Suhang Wang , Jiliang Tang and Huan Liu- "Fake News Detection using Data Mining Perspective".
[3] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
[4] H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383
[5] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279
[6] Scikit-Learn- Machine Learning In Python.

[7] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake News Detection on Social Media: A Data Mining Perspective” arXiv:1708.01967v3 [cs.SI], 3 Sep 2017
[8] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, “Fake News Detection on Social Media: A Data Mining Perspective” arXiv:1708.01967v3 [cs.SI], 3 Sep 2017
[9] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
[10]Fake news websites. (n.d.) Wikipedia. [Online]. “:?><LkmnjbhgvcfAvailable:https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017
[11] Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2017, pp. 208-215.

ABOUT THE AUTHORS

Ms. Aarti Chugh, Department of Computer Science, Amity University, Gurugram, Haryana, India

Dr. Yojana Arora, Department of Computer Science, Amity University, Gurugram, Haryana, India
.

Jaivardhan singh, Department of Computer Science, Amity University, Gurugram, Haryana, India

Shobhit, Department of Computer Science, Amity University, Gurugram, Haryana, India

Ronak Rathi, Department of Computer Science, Amity University, Gurugram, Haryana, India