IEEE-SEM, Volume 11, Issue 4, April-2023
ISSN 2320-9151

Anomalies Detection in Social Media News Using Machine Learning Approach

Daniel Terhemen Tuleun [a], Nazarov Alexey[b] Saugatdeep Ranabhat [c]

[a] Moscow Institute of Physics and Technology (MIPT), Moscow, Russia, tuleun.t@phystech.edu

OCID 0000-0002-6982-2710.

[b]Federal Research Center Computer Science and Control of Russian Academy of Sciences, Moscow, Russia, a.nazarov06@bk.ru. https://orcid.org/0000-0002-0497-0296.

[c] Moscow Institute of Physics and Technology (MIPT), Moscow, Russia, ranabkhat.s@phystech.edu

Abstract

Social media is a strong tool for discussing crucial issues such as politics and other related matters but if not properly handled can cause problem in the society and also it may have a negative impact on both the society and its economy. The extensive use of social media has both potential positive and negative effects on culture, business, and politics around the world. Social media coverage of crisis events, for instance, may be used by authorities to manage disasters effectively or by malicious parties to spread rumors and false information for financial or political gain. Given the adverse effects of fake news on social media, it is crucial to identify false information, keep it under control, and stop it from spreading. This study uses textual information that passes through search engines to collect and analyze potential false or misleading content. By using a real-world dataset associated with politics and other world news to find the best Machine learning approach that can work for detecting unreliable news from the real news. In the same vine we tried to bridge the gap in the literature by deploying some powerful Algorithms which are not commonly used by most researchers. All our algorithms performed excellent with high accuracy. In order to get the accurate performance as well as the prediction of our models, a confusion matrix was used to statistically analyze the result and finally we arrived at a conclusion that, out of the several algorithms we used for the task, passive aggressive classifier come up with the highest accuracy of 99%, showing that our accuracy outperformed all the previous research in this area and can be used for the purpose of anomalies detection of news on social media ad any task of this kind.

Keywords: Anomalies; Clickbait; Fake news; Social-Media; Machine learning

I. Introduction

With the continuous expansion of network scale and the rapid development of network applications, network security is becoming more and more important. Therefore, network anomaly detection has become an important research topic. Malicious attacks, node or link disconnection, among other things, are examples of network anomalies. Current research has demonstrated that any anomaly will result in an abnormal change in traffic volume. Thus, it is possible to discover network anomalies by keeping an eye on variations in traffic volume in a network. The security of computer systems and networks against attacks is currently subject to a number of threats and weaknesses. Along with the explosive growth of the Internet and the continued dramatic increase in all wireless services, the number and impact of attacks has been increasing. Recent well-publicized denial of service attacks against several well-known

web portals and numerous other such occurrences serve as evidence for this. The quantity of computer systems and their vulnerabilities has increased, while the level of technical attack knowledge required to carry out an attack has decreased due to the widespread availability of such knowledge on websites around the world. Computer network problems are discovered as traffic anomalies that they produce. An anomaly is typically characterized as something that defies logic. For example, a faulty switch may cause unexpected traffic in another section of the network, or new error codes may surface when a service is unavailable. The foundation of network troubleshooting is network anomalies. Another aspect of anomalies is the aspect that has to do with altering the news from its original content by either adding the sugar-coated words to it or removing some vital information from the original content, which is sometimes regarded to as fake news. Fake news spread faster and many people are always eager to listen and pay more attention to it than the real news because, it is spread with a target to destabilizes or turn people attention away from the original news and this is going to be our area of focus in this research.

Fake news as defined by Paskin (2018: 254), is "specific news articles that emerge on mainstream media online or offline, social media, or even both, and have no factual foundation but are presented as facts rather than satire." The COVID-19 pandemic served as a specific illustration of the need to combat false information. Social media platforms are increasing their use of digital tools for detecting fake news and educating users on how to recognize it. As of the time of writing (Sparks and Frishberg 2020), Facebook uses machine learning algorithms to identify sensational or false claims made in advertisements for alternative treatments. They also move potentially fake news articles lower in the news feed and give users advice on how to spot fake news on their own. Instagram directs anyone looking for information on the virus to a special message with credible information, and Twitter makes sure that searches on the virus lead to credible articles (Marr, 2020).

The fact that there are numerous methods for detecting fake news makes these measures possible. Platforms using machine learning, for instance, use fake news from the largest media outlets to develop algorithms for spotting fake news. Some approaches detect fake news by comparing the release time of the article to timelines of spreading the article as well as where the story spread.

II. Related Work

There are many available approaches to help the public to identify fake news and this research aims to enhance understanding of these by categorizing these approaches Fake news is not a new concept. Before the era of digital technology, it was spread through mainly yellow journalism with focus on sensational news such as crime, gossip, disasters and satirical news Prior to the information on COVID-19 pandermic, machine learning, and natural language processing have played an essential role in fighting misinformation and fake news (Braşoveanu & Andonie, 2019). We think the best course of action in this conflict is to develop fresh solutions that combine the strengths of both disciplines. There is a range of useful techniques and algorithms in the literature that illustrate both machine learning algorithms and natural language processing approaches, either separately or in combination. Here, we discuss the history and ideas that served as the foundation for this paper's methodology. In the early 2000 s, (Soon et al., 2001) claimed that training a Machine learning algorithm with specific linguistic features holds a promise in classifying text in general. According to the authors, their algorithm was the first learning-based system that was trained using bigram characteristics to produce outcomes that were on par with those of non-learning techniques (Mackey et al., 2020) where natural language processing and machine learning were combined in an effort to find potential false information on social media. The method found keywords connected to the epidemic and possible marketing. The authors used a deep learning system to analyze millions of social media postings and found a lot of questionable and unreliable products. (Fei Liu et al., 2008) presented a “survey-like” paper to demonstrate the various applications of combining both natural language processing and machine learning. This specifically covered the process of employing word features (bigrams) to train algorithms. Bigrams, which appear in texts as a pair of words (e.g., global pandemic). They offer more useful and complex textual properties than their simple counterparts of single high-frequency words. (Aphiwongsophon & Chongstitvatana, 2018) demonstrated how famous ML algorithms (e.g., Naïve Bayes, and Support

Vector Machines can be used to detect fake news. Their results showed promise with an accuracy of 96% or better. Following a similar path, H. (Ahmed et al., 2017) additionally employed a traditional support vector machine variant, or SVM, but trained the algorithms using n-gram data. The accuracy of their algorithm was lower than the previous methods (92%). The authors claimed that training the algorithm with n-grams was superior than using features with high frequency but no relevance to the dataset's context in terms of feature quality. Another interesting approach was employed by (Conroy et al., 2015) who else identified bogus news by using machine learning to detect deception. The strategy included network analysis for networks of connected data, machine learning, and linguistic aspects (such as n-grams). According to the authors, classification tasks for identifying false news have demonstrated high accuracy using both language and network analysis methods. Following their research, the writers offered the following suggestions:

1. achieving maximum performance requires deeper linguistic analysis and;
2. the utilization of linked data and a corresponding format will assist in achieving up-to-date fact checking.

(Oyebode & Orji, 2019) analysed public sentiments expressed towards two popular candidates in the Nigerian presidential election using lexicon-based and supervised machine learning (ML) approaches. their extended lexicon-based approach, VADEREXT, outperformed the other two approaches (i.e., VADER and TextBlob) with an overall F1 score of 76.3%. Also, the five ML models they built for the experiments surpassed the chance baseline, with LR achieving the best F1 score. VADER-EXT also achieved a better overall precision (81.6%) than LR. They also conducted thematic analysis on both positive and negative posts to further understand and reveal public opinions about each candidate by categorizing the posts into themes. Similarly, (Tumasjan et al., 2010) performed three research studies within the context of 2009 German federal election. By gathering tweets that either name the six political parties or well-known politicians in those parties, they first looked into whether Twitter actually fosters political discussion. Second, they analyzed whether tweets reflect political opinions expressed offline. Finally, they analysed whether volume of tweets reflects the popularity of parties in the real world and predicts election results. Their findings validate the popular belief that social media provides a platform for discussing political issues, and that social messages strongly reflect offline sentiments. (Razzaq et al., 2014) also analysed and predicted Pakistan general election using public sentiments expressed towards political parties on social media. They applied supervised machine learning techniques in classifying tweets into positive, negative, or neutral sentiments. They compared the average accuracies of several ML algorithms, including Naïve Bayes and SVM. Naïve Bayes performed best with an average accuracy of 70% for binary classification and about 55% for multiclass classification. Finally, (Asghar et al., 2014) applied the lexicon-based approach in their sentiment analysis tasks. They experimented with multiple lexicons and combined some of them top boost coverage. Due to its greater coverage, the hybrid lexicon with the largest size (labeled Hybrid-1) greatly improved its binary sentiment classification (i.e. positive versus negative) findings.

A. Limitations of Related Studies

The above introduction explains that the related methods motivate the subject and presents the current state of the art. It is clear that both machine learning and text mining present the corner stones for text classification and anomaly detection. However, regardless of the underlying algorithmic classification method (naïve Bayes, support vector machines), they were all trained from a static set of textual features, such as bigrams. Once the featured were derived, there has been no further work on how the features are related to each other to tell a much bigger story. Our network training model, however, connects the features in the way that the bigrams are naturally connected in the text. This offers the following advantages

1. It makes the model extensible by new datasets without doing the entire training;
2. A network model allows pruning (i.e., getting rid of the noise) using inherent centrality measures (degree, betweenness, closeness, etc.);
3. If necessary, a network model allows multi-label classification by applying network clustering techniques.

B. The Evolution of Fake News and Fake News Detection

Fake news is not a new concept. Before the era of digital technology, it was spread through mainly yellow journalism with focus on sensational news such as crime, gossip, disasters and satirical news (Stein-Smith et al., 2017). The prevalence of fake news relates to the availability of mass media digital tools (Gesellschaft, 2019). Since anyone can publish articles via digital media platforms, online news articles include well researched pieces but also opinion-based arguments or simply false information (Castelo et al., 2019). There is no custodian of credibility standards for information on these platforms making the spread of fake news possible. To make things worse, it is by no means straightforward telling the difference between real news and semi-true or false news (Pérez et al., 2017) The nature of social media makes it easy to spread fake news, as a user potentially sends fake news articles to friends, who then send it again to their friends and so on. Comments on fake news sometimes fuel its 'credibility' which can lead to rapid sharing resulting in further fake news (Albright, 2017). Social bots are also responsible for the spreading of fake news. Bots are sometimes used to target super-users by adding replies and mentions to posts. Humans are manipulated through these actions to share the fake news articles (Shao et al., 2018). Clickbait is another tool encouraging the spread of fake news. Clickbait is an advertising tool used to get the attention of users. Sensational headlines or news are often used as clickbait that navigate the user to advertisements. More clicks on the advert means more money (Chen et al. 2015a). Fortunately, tools have been developed for detecting fake news. For example, a tool has been developed to identify fake news that spreads through social media through examining lexical choices that appear in headlines and other intense language structures (Chen et al., 2015b). Another tool, developed to identify fake news on Twitter, has a component called the Twitter Crawler which collects and stores tweets in a database (Atodiresei et al., 2018). When a Twitter user wants to check the accuracy of the news found they can copy a link into this application after which the link will be processed for fake news detection. This process is built on an algorithm called the NER (Named Entity Recognition.

There are many available approaches to help the public to identify fake news and this research aims to enhance understanding of these by categorizing these approaches as found in existing literature.

III. Methodology

In this work we consider the anomalies detection in news on social media. Many datasets were collected and further divided in to training, testing and Visualization after that, many Machine learning (algorithms) such as Decision tree, Naïve Bayes, Passive Aggressive, Random Forest, XGBOOST Classifier and Logistics Regression were used in training, testing and prediction of the dataset. When the data was collected, after data cleaning, pre-processing, and wrangling, the first step we did was to feed it to an outstanding model and of course, get output in probabilities. A confusion matrix was used to measure the effectiveness, to better the effectiveness and the performance of the model. Confusion Matrix is a performance measurement for machine learning classification. The figure1 bellow shows details on the operation on the dataset, from the collection, training, testing, visualization and prediction.

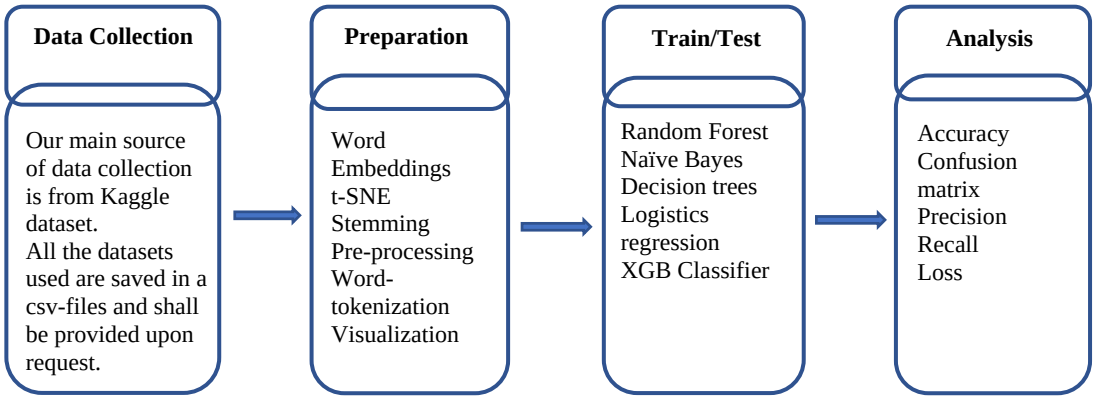

Fig.1 Data collection training, testing and Analysis

A. Dataset Collection

As discussed earlier, the first process for building any Machine learning model is the collection or mining of data. For this project, Kaggle Dataset is identified as the main source of data for all the Machine learning models to be compared. A Python program language is used to work on the data set to detect real News from fake News collect texts from Kaggle. The main dataset here is that of 2016 US Presidential election. All the datasets used in this work were saved and can be provided to anyone for any reasonable request.

B. Performance measurement

i. Accuracy: the percentage of events that were successfully predicted compared with all the predictions. Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ (1)

ii. Precision: all true positive divided by all positive predictions, presented as follows =$\frac{TP}{TP+FP}$ (2)

iii. Recall: true positives divided by positive results. This pattern indicates that out of potential positives, how many were found by the model? = $\frac{TP}{TP+FN}$ (3)

iv. F1-Scores = 2 x $\frac{Precision\ x\ Recall}{Precision+\ Recall}$ (4)

IV. Result

In this work we tried to detect anomalies on Social Media (Fake news detection) using different Machine learning algorithms. Two different Dataset were obtained from Kaggle for this purpose. The datasets were divided into training and testing, 80% for training and 20% for testing. A pre-processing and stemming methods were used on the variant forms of words to reduce them to common form. In Data Pre-processing following steps are followed: Firstly, all the sequences except English characters are removed from the string. Next, to avoid false predictions or ambiguity with upper and lowercase, all the characters in strings are converted to lowercase. Furthermore, all the sentences are tokenized into words. To facilitate fast processing, stemming is applied to the tokenized words. And finally, words are joined together and stored in the corpus. Different Algorithms were built on the datasets to test the present of false or unreliable news. Logistics Regression, Naïve Bayes, Decision Tree, Random Forest, Passive Aggressive Classifier and XGB Classifier. The results are giving by accuracy and confusion matrix. Logistic Regression scored 98%, Naïve Bayes scored 67%, Decision Tree scored 90%, Random Forest scored 90%, and Passive Aggressive Classifier scored 99% while XGB Classifier scored 91%. A word cloud was built for a better visualization on the dataset for the purpose of the univariate analysis. A word cloud is a visualization approach for text data where the most common term is presented in the most considerable font size. Bivariate Analysis, Bigram and Trigram were also used.

S/No	Model Name	Precision	Recall	F1-Score	Accuracy
1	LOGISTICS REGRESSION	98	98	98	98
2	NAÏVE BAYES	67	67	67	67
3	DECISION TREE	90	90	90	90
4	RANDOM FOREST	91	90	90	90
5	XGB CLASSIFIER	92	91	91	91

6	PASSIVE AGGRESSIVE	99	99	99	99

Table 1. Classification Report of the Models.

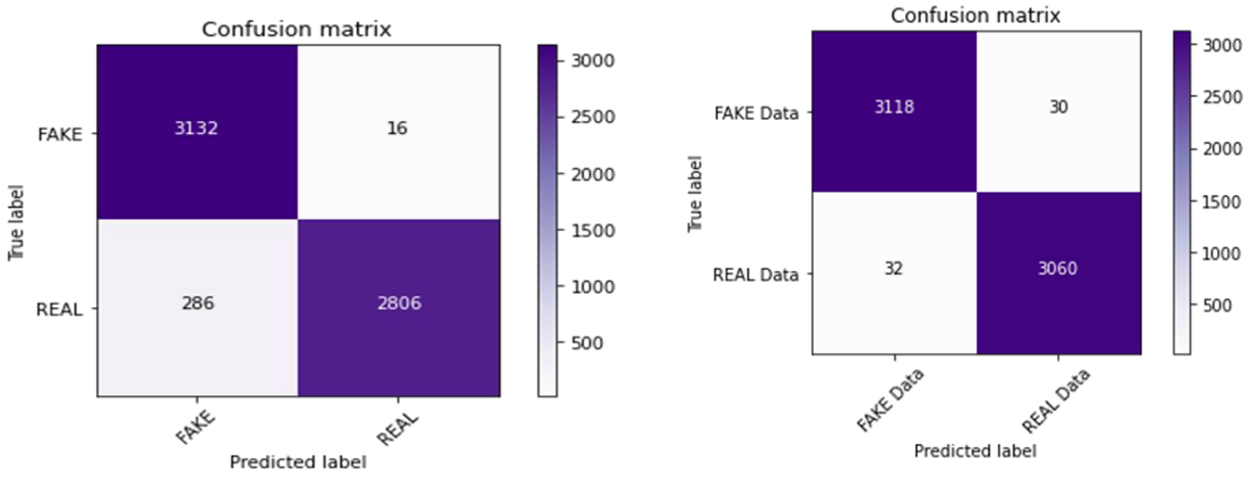

Figure (2a, b). Confusion matrix for Multinomial Naïve Bayes (left) and Passive Aggressive (right) respectively

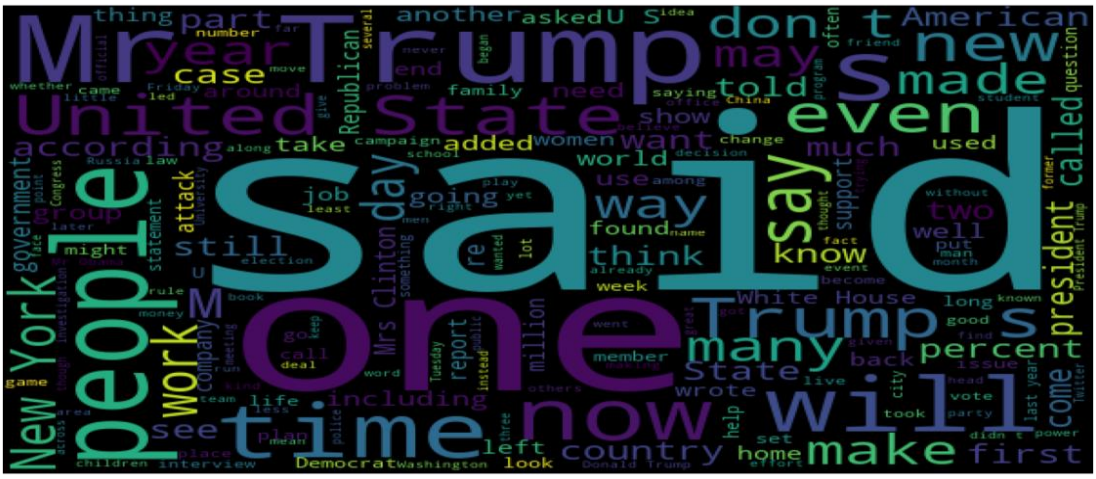

Figure 3a. Word Cloud for Donald Trump.

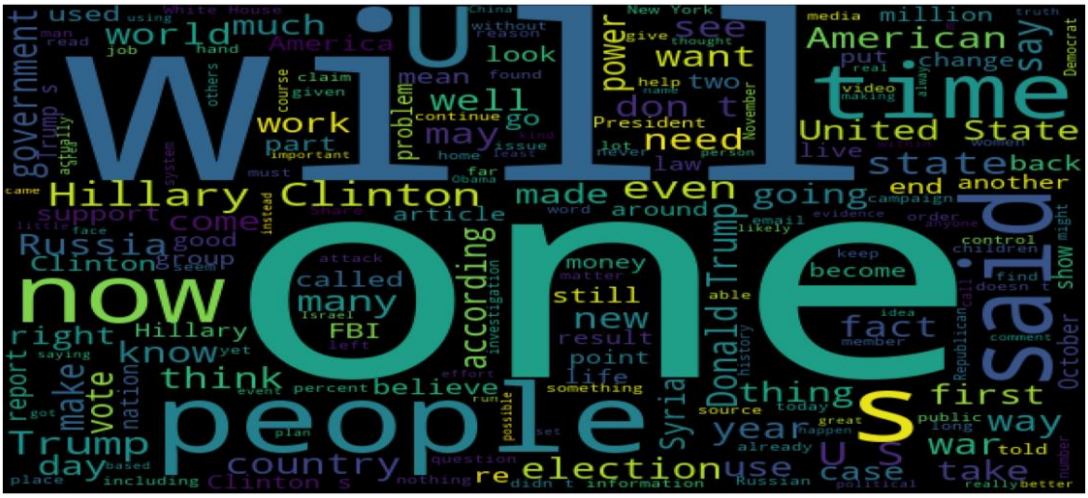

Figure 3b. Word Cloud for Hillary Clinton

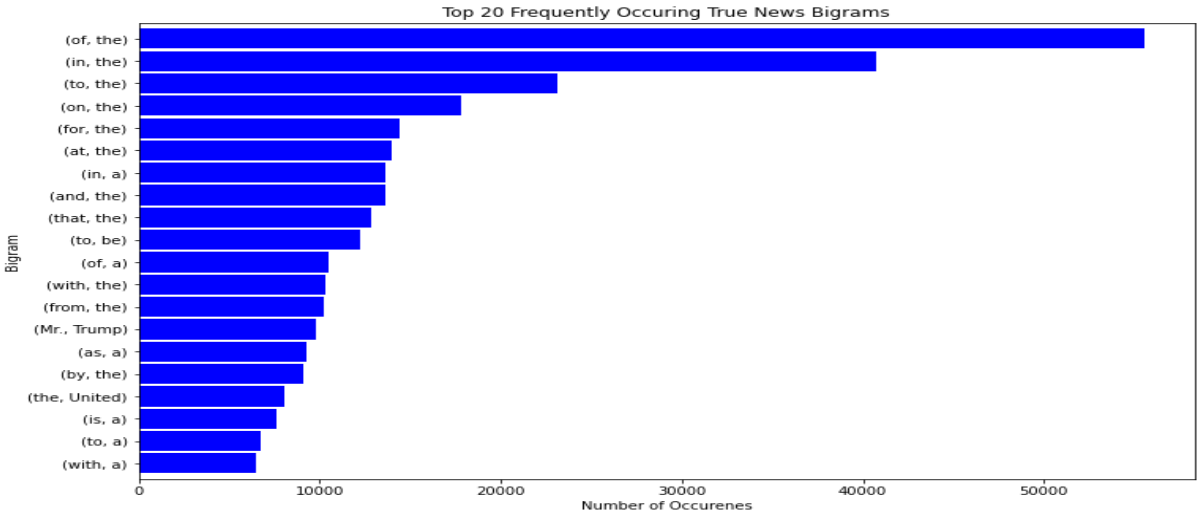

Figure 4a. N= 2 Reliable news

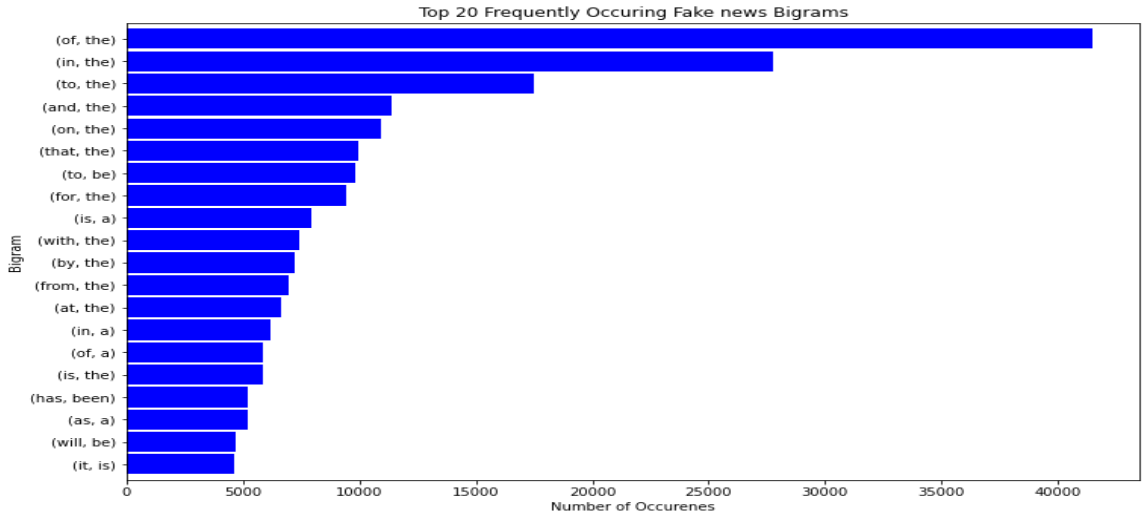

Figure 4b. (N=2 Unreliable)

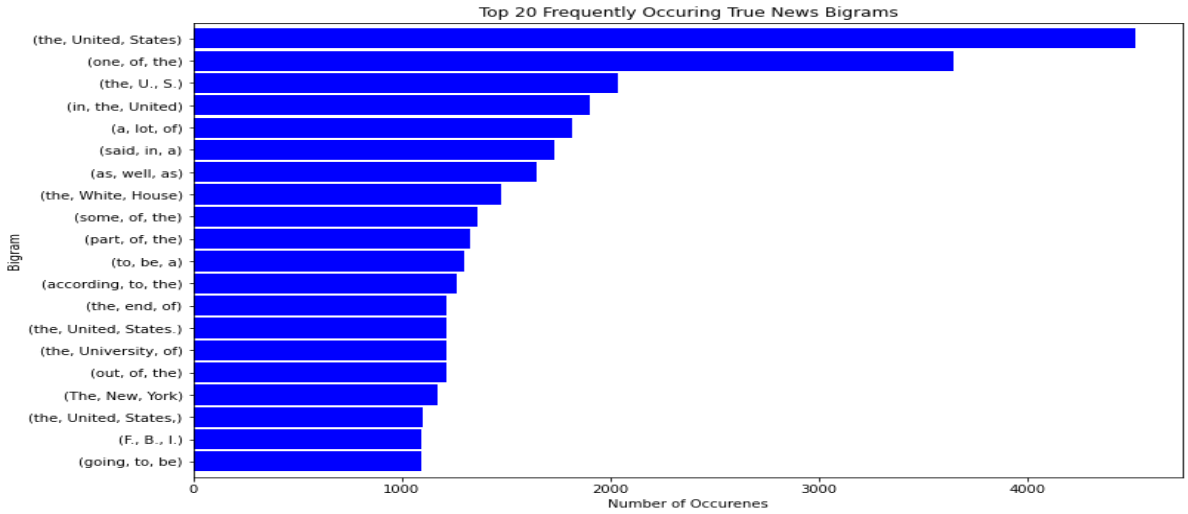

Figure 5a. (N=3 Reliable news)

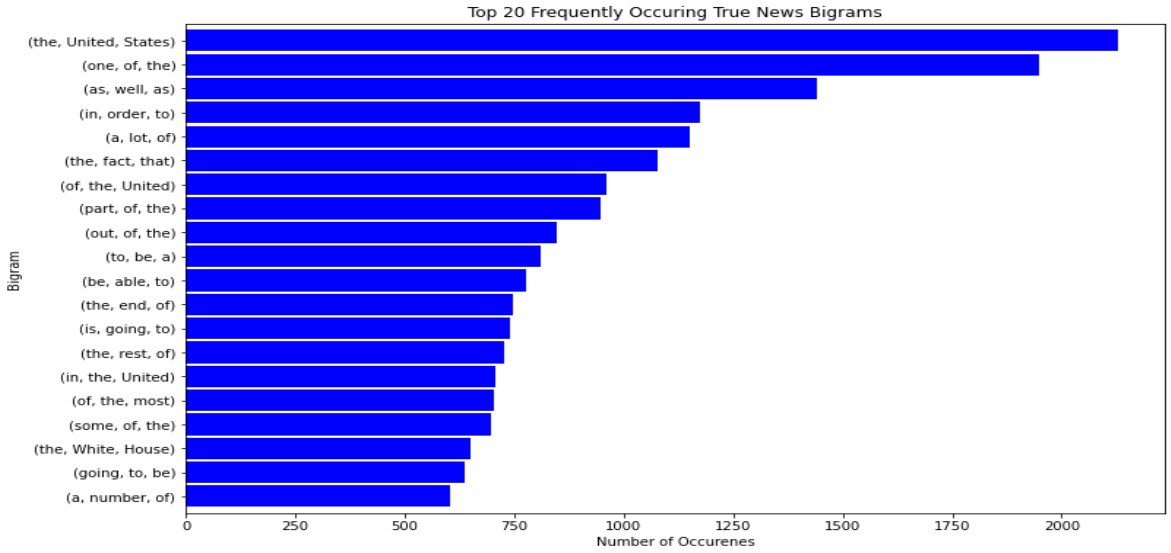

Figure 5b. (N=3 Unreliable news)

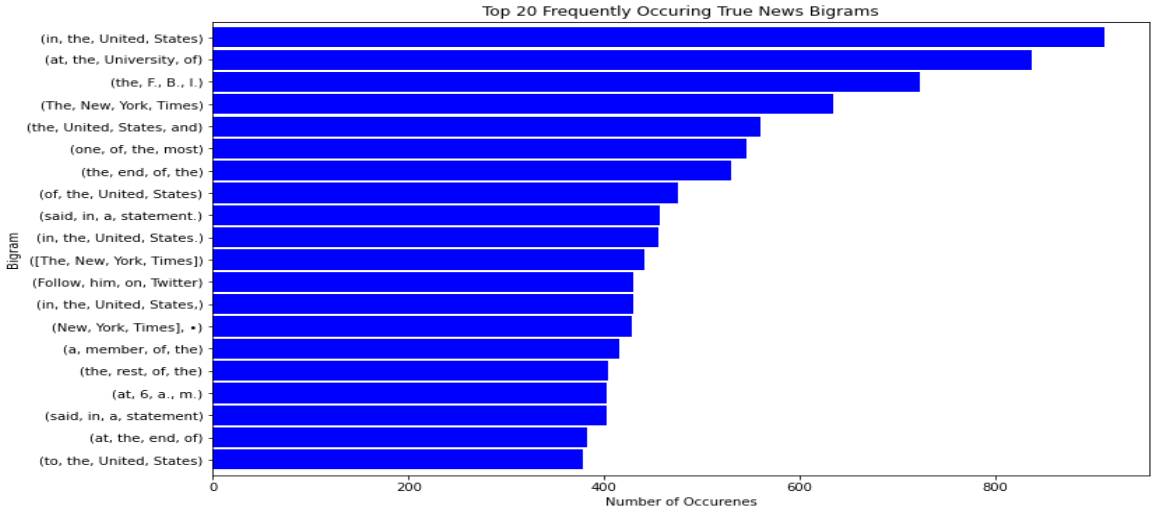

Figure 6a. (N=4 Reliable news)

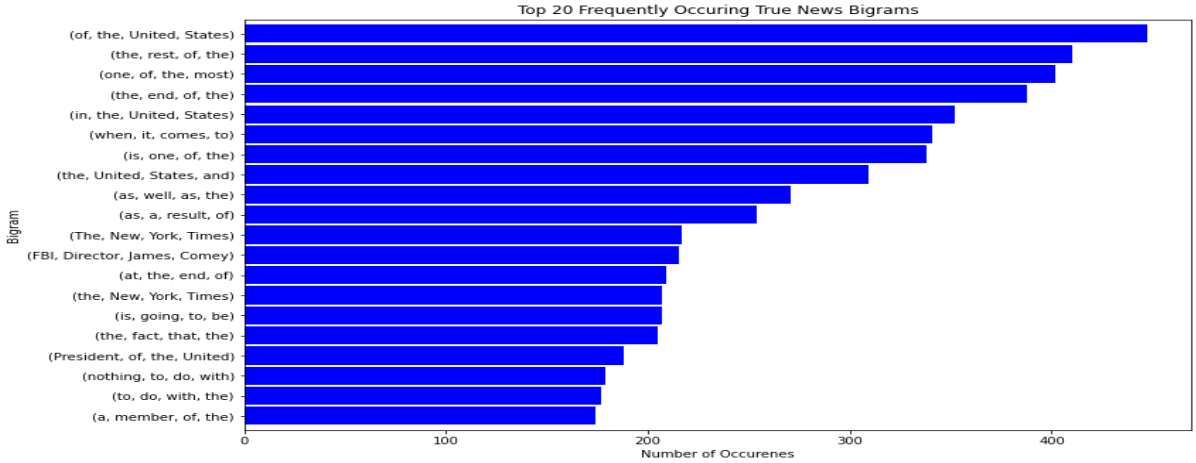

Figure 6b. (N=4 Unreliable news)

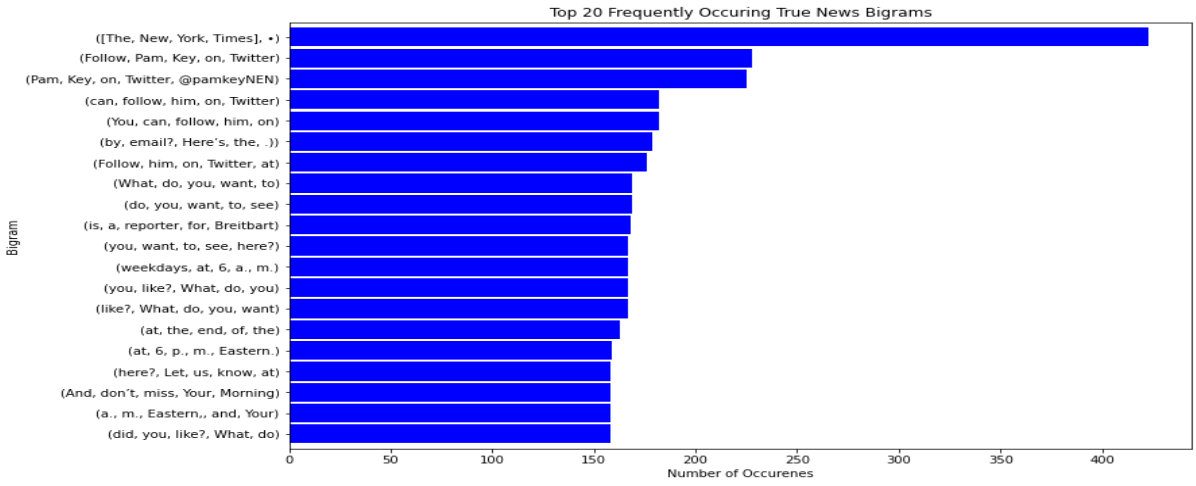

Figure 7a. (N=5 Reliable news)

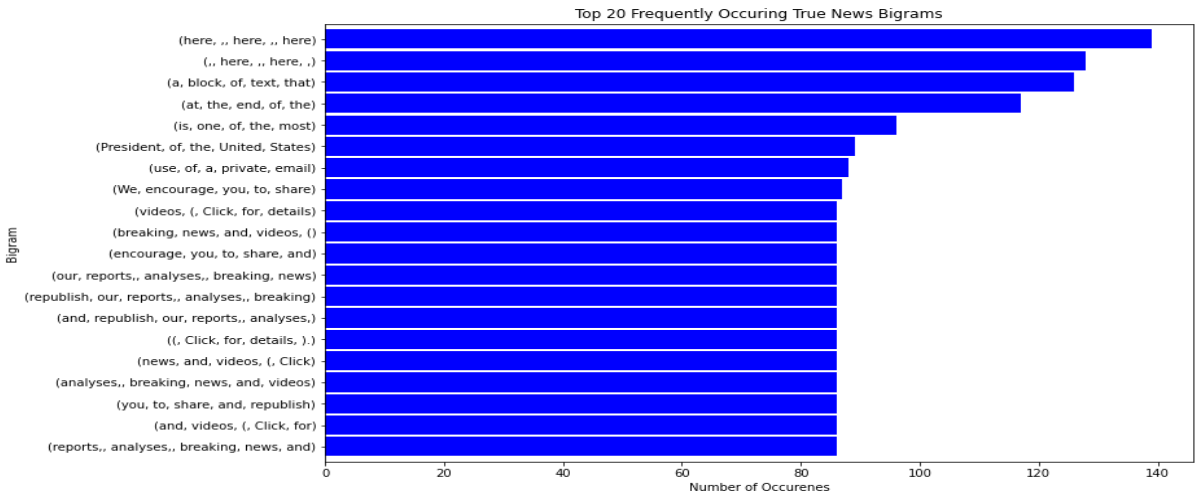

Figure 7b. (N=5 Unreliable news)

V. Conclusion and Future Research

Anomalies detection in news on social media is studied. In this work we tried to detect unreliable news from the reliable news using different Machine learning algorithms. The Dataset used in this work is that of the 2016 US presidential election which is available on Kaggle. The datasets were divided into training and testing, 80% for training and 20% for testing. A pre-processing and stemming methods were used on the variant forms of words to reduce them to common form. Furthermore, the following algorithms were used to test the present of false or unreliable news. Logistics Regression, Naïve Bayes, Decision Tree, Random Forest Passive Aggressive Classifier and XGB Classifier. Interesting results were obtained and are giving by accuracy and confusion matrix. Logistic Regression scored 98%, Naïve Bayes scored 67%, Decision Tree scored 90%, Random Forest scored 90% Passive Aggressive Classifier scored 99% while XGB Classifier scored 91%. A word cloud was built for the two candidates Donald Trump and Hillary Clinton for the purpose of the univariate analysis. A word cloud is a visualization approach for text data where the most common term is presented in the most considerable font size. Bivariate Analysis, Bigram and Trigram were also used. From our visualization of both word cloud, Bivariate analysis and trigram one could easily see that the result is in favour of Donald Trump as the winner of the 2016 US presidential election going by our model, which is contrary to the news that was spreading round that Hillary Clinton was wining. Our results match the final result that was later announce by the electoral body which build more confidence in our algorithms. Furthermore, this method can be used to test and predict results of this nature from other country. A comparative study of the result revealed that, Passive Aggressive classifier scored the highest accuracy of 99% with a very low false alarm followed by logistics regression with the accuracy of 98%. While Naïve Bayes come up with the least accuracy of 67%.

For further research I would like to improve the detection result by implementing more strong algorithms in order to select the best algorithm for this purpose that will not just give a better accuracy but a 100% accuracy with no false alarm. Furthermore, we are planning to build some model prediction in the future that can detect rigging in any election,

Acknowledgment

The authors are grateful to their research advisor Ilia Mikhailovich Voronkov and Alexey N. Nazarov for their contributions during the lab work and the team responsible for vetting the manuscript before the final publication. Also, my special appreciation goes to my mother Rahel Mbakunden Tuleun for her moral support always.

References

Oyebode, O., & Orji, R. (2019). Social media and sentiment analysis: The Nigeria Presidential election 2019. *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. https://doi.org/10.1109/iemcon.2019.8936139

Chen, C.-H., Huang, W.-T., Tan, T.-H., Chang, C.-C., & Chang, Y.-J. (2015). Using K-nearest neighbor classification to diagnose abnormal lung sounds. *Sensors*, *15*(6), 13132–13158. https://doi.org/10.3390/s150613132

Stein-Smith, K. (2017) Librarians, Infor-mation Literacy and Fake News: Helping Students to Tell the Difference between Al-ternative Facts and the Real News. Strategic Library, No. 37, 1-23.

Albright, J. (2017). Welcome to the era of fake news. *Media and Communication*, *5*(2), 87–89. https://doi.org/10.17645/mac.v5i2.977

Atodiresei, C.-S., Tănăselea, A., & Iftene, A. (2018). Identifying fake news and fake users on Twitter. *Procedia Computer Science*, *126*, 451–461. https://doi.org/10.1016/j.procs.2018.07.279

Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., & Freire, J. (2019). A topic-agnostic approach for identifying fake news pages. *Companion Proceedings of The 2019 World Wide Web Conference*. https://doi.org/10.1145/3308560.3316739

Asghar, Dr. Muhammad & Kundi, Fazal & Khan, Aurangzeb & Ahmad, Shakeel. (2014). Lexicon-Based Sentiment Analysis in the Social Web. journal of basic and applied scietific research. 4. 238-248.

Marr, B. (2020, March 27). *Coronavirus fake news: How facebook, Twitter, and Instagram are tackling the problem*. Forbes. Retrieved January 5, 2023, from https://www.forbes.com/sites/bernardmarr/2020/03/27/finding-the-truth-about-covid-19-how-facebook-twitter-and-instagram-are-tackling-fake-news/?sh=2c340fdb1977

Paskin, D. (2018). Real or Fake News: Who Knows? *Social media and society*, *7*, 252-273.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178–185. https://doi.org/10.1609/icwsm.v4i1.14009

Software that can automatically detect fake news. Fraunhofer. (2019, February 11). Retrieved Jan 5, 2023, from https://www.fraunhofer.de/en/press/research-news/2019/february/software-that-can-automatically-detect-fake-news.html

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by Social Bots. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06930-7

Sparks, H., & Frishberg, H. (2020, March 26). *Facebook gives step-by-step instructions on how to spot fake news*. New York Post. Retrieved April 6, 2023, from https://nypost.com/2020/03/26/facebook-gives-step-by-step-instructions-on-how-to-spot-fake-news/

Chen, C.-H., Huang, W.-T., Tan, T.-H., Chang, C.-C., & Chang, Y.-J. (2015). Using K-nearest neighbor classification to diagnose abnormal lung sounds. *Sensors*, *15*(6), 13132–13158. https://doi.org/10.3390/s150613132

Razzaq, M. A., Qamar, A. M., & Hafiz Syed Muhammad Bilal. (2014). Prediction and analysis of Pakistan election 2013 based on sentiment analysis. *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. https://doi.org/10.1109/asonam.2014.6921662

Braşoveanu, A. M., & Andonie, R. (2019). Semantic Fake News Detection: A machine learning perspective. *Advances in Computational Intelligence*, 656–667. https://doi.org/10.1007/978-3-030-20521-8_54

Soon, W. M., Ng, H. T., & Lim, D. C. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521–544. https://doi.org/10.1162/089120101753342653

Mackey, T. K., Li, J., Purushothaman, V., Nali, M., Shah, N., Bardier, C., Cai, M., & Liang, B. (2020). Big Data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram. *JMIR Public Health and Surveillance*, *6*(3). https://doi.org/10.2196/20794

Fei Liu, Feifan Liu, & Yang Liu. (2008). Automatic keyword extraction for the meeting corpus using supervised approach and Bigram expansion. *2008 IEEE Spoken Language Technology Workshop*. https://doi.org/10.1109/slt.2008.4777870

Aphiwongsophon, S., & Chongstitvatana, P. (2018). Detecting fake news with Machine Learning Method. *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. https://doi.org/10.1109/ecticon.2018.8620051

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using N-gram analysis and Machine Learning Techniques. *Lecture Notes in Computer Science*, 127–138. https://doi.org/10.1007/978-3-319-69155-8_9

Conroy, Nadia & Rubin, Victoria & Chen, Yimin. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology. 52. 1-4. 10.1002/pra2.2015.145052010082.