FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING

PROJECT REPORT
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF BACHELOR OF TECHNOLOGY
IN
DEPARTMENT OF INFORMATION TECHNOLOGY

BY

NAME	Registration No:	Roll No:
SHIVAM GHOSH	161420110110	14200216011
SHINJA GHOSH	161420110109	14200216012
SHILPI KUNDU	161420110108	14200216013
SATYAM SETH	161420110104	14200216017

Under The Supervision
of
<SUROJIT DAS & INDRAJIT DAS
ASSISTANT PROFESSOR, IT DEPARTMENT>

DEPARTMENT OF INFORMATION TECHNOLOGY

MEGHNAD SAHA INSTITUTE OF TECHNOLOGY

Techno Complex, Madurdaha, Beside NRI Complex, post-Uchhepota, Kolkata-700150

Affiliated by

MAULANA ABUL KALAM AZAD
UNIVERSITY OF TECHNOLOGY,
WEST BENGAL

JUNE 2020

CANDIDATE’S DECLARATION

I hereby declare that the work which is being presented in the project entitled fake news detection using natural language processing in fulfillment of requirements for the award of degree of B.Tech. in IT, submitted in the Department of Information Technology at **MEGHNAD SAHA INSTITUTE OF TECHNOLOLY** under **MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL** is an authentic record of our own work carried out during Session 2019-2020 under the supervision of **Surajit Das and Indrajit Das.** The matter presented in this project has not been submitted by us in any other University / Institute for any award.

Shilpi Kundu

Shivam Ghosh

Shinja Ghosh

Satyam Seth

Signature of the Students
With Date

Meghnad Saha Institute of Technology

Nazirabad, P.O:Utchepota, Via Sonarpur, Kolkata 700 150

CERTIFICATE

This is to certify that the Project entitled fake news detection using natural language processing is being submitted by Shilpi Kundu, Shivam Ghosh, Shinja Ghosh and Satyam Seth in partial fulfillment of the requirement for the award of the degree of B.Tech.in Information Technology to the Department of Information Technology, Meghnad Saha Institute of Technology, Kolkata, is a record of bonafied work carried out by him under my guidance and supervision from June, 2019 to July, 2020.

The results presented in this thesis have been verified and are found to be satisfactory. The results embodied in this thesis have not been submitted to any other University for the award of any other degree or diploma.

Surajit Das and Indrajit Das
Assistant Professor
Information Technology
MeghnadSaha Institute of Technolgy
Kolkata-700150

(Signature of the Project Guide)

Date:

Place:

Meghnad Saha Institute of Technology
Nazirabad, P.O. :Utchepota, Via Sonarpur, Kolkata 700 150

CERTIFICATE OF APPROVAL

The foregoing project entitle fake news detection using natural language processing is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as prerequisite for the degree for which it has been submitted. It is to be understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Board of Examiners

Project Guide

Head of the Department

Date:

Place:

[Departmental & College Seal]

Acknowledgement

Life as a researcher is a zig-zag way and has many ups and downs, but I overcome from them because of the support and faith of many individuals.

First and foremost, I would like to express my sincere gratitude to **Mr. Surajit Das and Mr. Indrajit Das** for providing me an opportunity to work with them. Your creativity encouraged me, your energy and dedication towards work motivates me.Your perfection always inspired me to do things perfectly. I feel motivated and encouraged every time I attend your meeting. Without your encouragement, inspiration and guidance this thesis / dissertation would not have materialized. Thank you so much for not only guiding me in this thesis / dissertation but for every knowledge you give me. This thesis / dissertation is dedicated to your creativity and enthusiasm of doing work.

I wish to thanks many people for their direct or indirect help in achieving this goal. Special thanks to my parents for their support and encouragement throughout my study, their faith in me ignite that spark to do anything for their happiness.

Last but not least I would like to thanks all my friends for being with me at each step when I need their support. This thesis never be successful without your support and love.

Shilpi Kundu

Shivam Ghosh

Shinja Ghosh

Satyam Seth

____

(Signature of the Students)

Abstract

Fake news is a phenomenon which is having a significant impact on our social life, in particular in the political world. Fake news detection is an emerging research area which is gaining interest but involved some challenges due to the limited amount of resources (i.e., datasets, published literature) available. We propose in this paper, a fake news detection model that use n-gram analysis and machine learning techniques. We investigate and compare two different features extraction techniques and five different machine learning classification techniques. Experimental evaluation yields the best performance using Term Frequency-Inverted Document Frequency (TF-IDF) as feature extraction technique.

Keywords: Online fake news, Text classification, Online social network security, Fake news detection, N-gram analysis.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1. INTRODUCTION

In the recent years, online content has been playing a significant role in swaying users' decisions and opinions. Opinions such as online reviews are the main source of information for e-commerce customers to help with gaining insight into the products they are planning to buy.

Recently it has become apparent that opinion spam does not only exist in product reviews and customers' feedback. In fact, fake news and misleading articles is another form of opinion spam, which has gained traction. Some of the biggest sources of spreading fake news or rumors are social media websites such as Google Plus, Facebook, Twitters, and other social media outlet [7].

Even though the problem of fake news is not a new issue, detecting fake news is believed to be a complex task given that humans tend to believe misleading information and the lack of control of the spread of fake content [8]. Fake news has been getting more attention in the last couple of years, especially since the US election in 2016. It is tough for humans to detect fake news. It can be argued that the only way for a person to manually identify fake news is to have a vast knowledge of the covered topic. Even with the knowledge, it is considerably hard to successfully identify if the information in the article is real or fake. The open nature of the web and social media in addition to the recent advance in computer science simplify the process of creating and spreading fake news. While it is easier to understand and trace the intention and the impact of fake reviews, the intention, and the impact of creating propaganda by spreading fake news cannot be measured or understood easily. For instance, it is clear that fake review affects the product owner, customer and online stores; on the other hand, it is not easy to identify the entities affected by the fake news. This is because identifying these entities require measuring the news propagation, which has shown to be complex and resource intensive [9]. Trend Micro, a cyber security company, analyzed hundreds of fake news services provider around the globe. They reported that it is effortless to purchase one of those services. In fact, according to the report, it is much cheaper for politicians and political parties to use those services to manipulate election outcomes and people opinions about certain topics [10, 11]. Detecting fake news is believed to be a complex task and much harder than detecting fake product reviews given that they spread easily using social media and word of mouth.

We present in this paper an n-gram features-based approach to detect fake news, which consists of using text analysis based on n-gram features and machine learning classification techniques. We study and compare five different supervised classification techniques, namely, Multinomial NB, Gradient Boosting, Random Classifier, Passive aggressive Classifier, Multinomial Classifier with Hyperparameter(alpha=1).

.

1.1. OBJECTIVE

The extensive spread of fake news can have a serious negative impact on individuals and society. The spread of fake news is a matter of concern as it manipulates the public opinions. During the American Presidential elections of 2016, it was estimated that over 1 million tweets are related to fake news “Pizza gate” by the end of the elections. The wide spread of fake news can have a huge negative impact on individuals and society as a whole. We aim to develop a model using machine learning and NLP techniques to determine whether a news is fake or real.

The major objectives of this project are:

1. Taking the help of N-gram modelling to develop a machine learning based model for accurately determining whether the given news is fake or authentic.
2. To get high accuracy to determine a news is fake or true.

1.2. DOMAIN DEFINATION

Fake news is a real menace as it can quickly spread panic among the public. It can also affect major world events, as was seen in the US Presidential Elections. With the flood of news arising from online content generators, as well as various formats and genres, it is impossible to verify news using traditional fact checkers and vetting. To tackle this problem of quick and accurate classification of news as fake or authentic, we provide a computational tool.

1.3. MOTIVATION OF RESEARCH

The extensive spread of fake news can have a serious negative impact on individuals and society. It has brought down the authenticity of news ecosystem as it is even more widely spread on social media than most popular authentic news. It is one of the biggest problems which has the ability to change opinions and influence decisions and interrupts the way in which people responds to real news.

The spread of fake news is a matter of concern as it manipulates the public opinions. During the American Presidential elections of 2016, it was estimated that over 1 million tweets are related to fake news “Pizza gate” by the end of the elections. It has political influence, can encourages mistrust in legitimate media outlet, influence financial markets, damage to individual’s reputation. The wide spread of fake news can have a huge negative impact on individuals and society as a whole During the American presidential elections of 2016, a survey revealed that many young men and teens in Veles were running hundreds of websites which published many false viral stories that supported Trump. This fake news influenced many people which affected the election results. This is just a small instance of how the spread of fake news can influence people.

Many organizations have come forward to stop the spread of fake news. E.g. Google app uses Artificial Intelligence to select stories and stop fake news. We aim to develop a model using machine learning and NLP techniques to determine whether a news is fake or real.

2. LITERATURE REVIEW

Research on fake news detection is still at an early stage, as this is a relatively recent phenomenon, at least regarding the interest raised by society. We review some of the published work in the following. In general, Fake news could be categorized into three groups. The first group is fake news, which is news that is completely fake and is made up by the writers of the articles. The second group is fake satire news, which is fake news whose main purpose is to provide humor to the readers. The third group is poorly written news articles, which have some degree of real news, but they are not entirely accurate. In short, it is news that uses, for example, quotes from political figures to report a fully fake story. Usually, this kind of news is designed to promote certain agenda or biased opinion [1].

Rubin et al. [2] discuss three types of fake news. Each is a representation of inaccurate or deceptive reporting. Furthermore, the authors weigh the different kinds of fake news and the pros and cons of using different text analytics and predictive modeling methods in detecting them. In this paper, they separated the fake news types into three groups:

– Serious fabrications are news not published in mainstream or participant media, yellow press or tabloids, which as such, will be harder to collect.

– Large-Scale hoaxes are creative and unique and often appear on multiple platforms. The authors argued that it may require methods beyond text analytics to detect this type of fake news.

– Humorous fake news, are intended by their writers to be entertaining, mocking, and even absurd. According to the authors, the nature of the style of this type of fake news could have an adverse effect on the effectiveness of text classification techniques. The authors argued that the latest advance in natural language processing (NLP) and deception detection could be helpful in detecting deceptive news. However, the lack of available corpora for predictive modeling is an important limiting factor in designing effective models to detect fake news.

Horne et al. [3] illustrated how obvious it is to distinguish between fake and honest articles. According to their observations, fake news titles have fewer stop-words and nouns, while having more nouns and verbs. They extracted different features grouped into three categories as follows:

– Complexity features calculate the complexity and readability of the text.

– Psychology features illustrate and measure the cognitive process and personal concerns underlying the writings, such as the number of emotion words and casual words.

– Stylistic features reflect the style of the writers and syntax of the text, such as the number of verbs and the number of nouns.

The aforementioned features were used to build a SVM classification model. The authors used a dataset consisting of real news from BuzzFeed and other news websites, and Burfoot and Baldwin's satire dataset [4] to test their model. When they compared real news against satire articles (humorous article), they achieved 91% accuracy. However, the accuracy dropped to 71% when predicting fake news against real news.

Wang et al. [5] introduced LIAR, a new dataset that can be used for automatic fake news detection. Thought LIAR is considerably bigger in size, unlike other data sets, this data set does not contain full articles, it contains 12800 manually labeled short statements from politicalFact.com.

Rubin et al. [6] proposed a model to identify satire and humor news articles. They examined and inspected 360 Satirical news articles in mainly four domains, namely, civics, science, business, and what they called "soft news" ('entertainment/gossip articles'). They proposed an SVM classification model using mainly five features developed based on their analysis of the satirical news. The five features are Absurdity, Humor, Grammar, Negative Affect, and Punctuation. Their highest precision of 90% was achieved using only three combinations of features which are Absurdity, Grammar, and Punctuation.

3. PROPOSED WORK

3.1 **N-gram Model**

N-gram modeling is a popular feature identification and analysis approach used in language modeling and Natural language processing fields. N-gram is a contiguous sequence of items with length n. It could be a sequence of words, bytes, syllables, or characters. The most used n-gram models in text categorization are word-based and character-based n-grams. In this work, we use word-based n-gram to represent the context of the document and generate features to classify the document. We develop a simple n-gram based classifier to differentiate between fake and honest news articles. The idea is to generate various sets of n-gram frequency profiles from the training data to represent fake and truthful news articles. We used several baseline n-gram features based on words and examined the effect of the n-gram length on the accuracy of different classification algorithms.

3.2 **Data Pre-processing**

Before representing the data using n-gram and vector-based model, the data need to be subjected to certain refinements like stop-word removal, tokenization, a lower casing, sentence segmentation, and punctuation removal. This will help us reduce the size of actual data by removing the irrelevant information that exists in the data. We created a generic processing function to remove punctuation and non-letter characters for each document; then we lowered the letter case in the document. In addition, an n-gram word-based tokenizer was created to slice the text based on the length of n. Stop Word Removal Stop words are insignificant words in a language that will create noise when used as features in text classification. These are words commonly used a lot in sentences to help connect thought or to assist in the sentence structure. Articles, prepositions and conjunctions and some pronouns are considered stop words. We removed common words such as, a, about, an, are, as, at, be, by, for, from, how, in, is, of, on, or, that, the, these, this, too, was, what, when, where, who, will, etc. Those words were removed from each document, and the processed documents were stored and passed on to the next step. Stemming After tokenizing the data, the next step is to transform the tokens into a standard form. Stemming simply is changing the words into their original form, and decreasing the number of word types or classes in the data. For example, the words “Running”, “Run” and “Runner” will be reduced to the word “run.” We use stemming to make classification faster and efficient. Furthermore, we use Porter stemmer, which is the most commonly used stemming algorithms due to its accuracy.

3.3 **Features Extraction**

One of the challenges of text categorization is learning from high dimensional data. There is a large number of terms, words, and phrases in documents that lead to a high computational burden for the learning process. Furthermore, irrelevant and redundant features can hurt the accuracy and performance of the classifiers. Thus, it is best to perform feature reduction to reduce the text feature size and avoid large feature space dimension. We studied in this research two different features selection methods, namely, Count Vectorizer (bag of words) and Term Frequency-Inverted Document Frequency (TF-IDF). These methods are described in the following.

The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. It is used to convert a collection of text documents to a vector of term/token counts. It also enables the pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.

The Term Frequency-Inverted Document Frequency (TF-IDF) is a weighting metric often used in information retrieval and natural language processing. It is a statistical metric used to measure how important a term is to a document in a dataset. A term importance increases with the number of times a word appears in the document; however, this is counteracted by the frequency of the word in the corpus. One of the main characteristics of IDF is it weights down the term frequency while scaling up the rare ones. For example, words such as “the” and “then” often appear in the text, and if we only use TF, terms such as these will dominate the frequency count. However, using IDF scales down the impact of these terms.

3.4 **Classification Process**

Figure 1 is a diagrammatic representation of the classification process. It starts with preprocessing the data set, by removing unnecessary characters and words from the data. N-gram features are extracted, and a features matrix is formed representing the documents involved. The last step in the classification process is to train the classifier. We investigated different classifiers to predict the class of the documents. We investigated specifically five different machine learning algorithms, namely, Multinomial NB, Gradient Boosting, Random Classifier, Passive aggressive Classifier, Multinomial Classifier with Hyperparameter(alpha=1). We used implementations of these classifiers from the Python Natural Language Toolkit (NLTK). We split the dataset into training and testing sets. For instance, in the experiments presented subsequently, we use 5-fold cross validation, so in each validation around 80% of the dataset is used for training and 20% for testing.

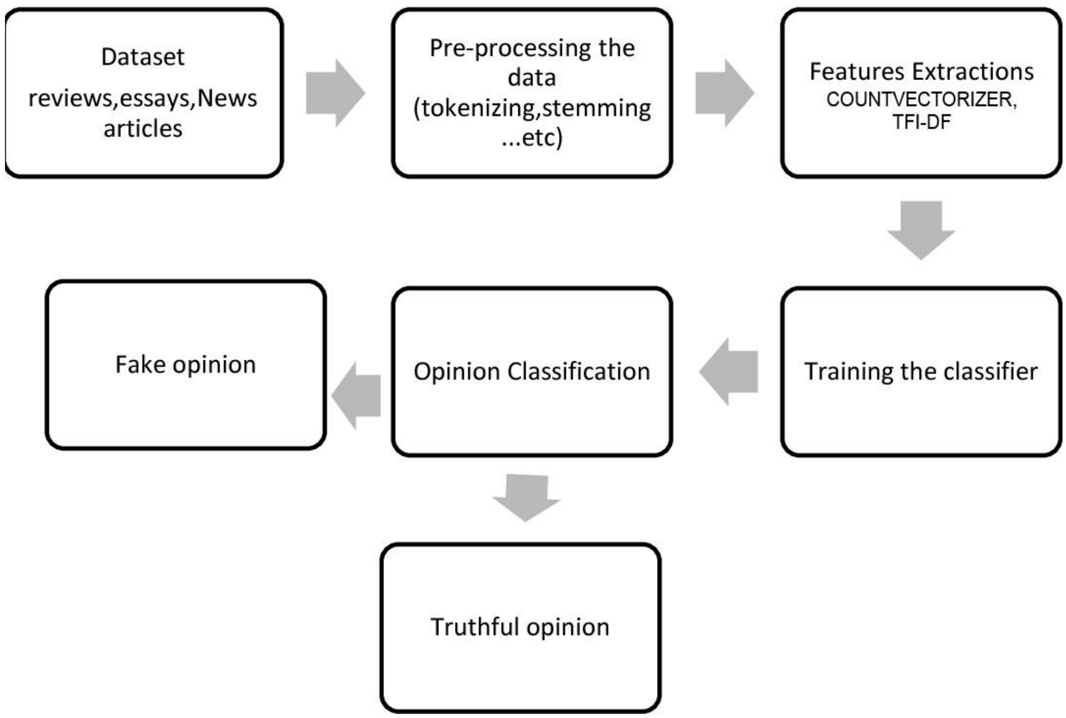


Fig. 1. Classification Process

4. EXPERIMENTS

4.1 DATASETS

We have used the fake news dataset from Kaggle.com to train our model. There are a total of around 20,800 data points, out of which 10413 articles are real and 10387 articles are fake initially. But after the data pre-processing, i.e., removing duplicates and null data points, there are about 20203 data points, out of which 10387 articles are real and 9816 articles are fake.

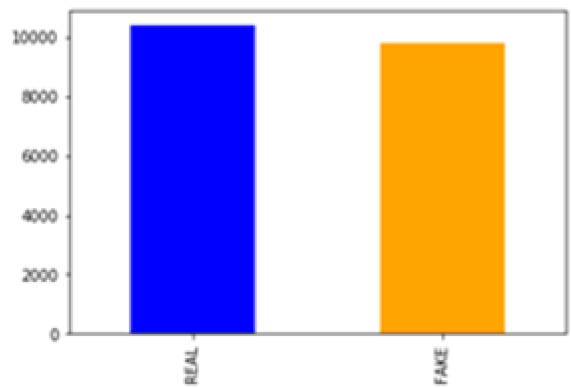


Fig. 2. Distribution of Datasets

4.2 EXPERIMENT ANALYSIS

We run the aforementioned machine learning algorithms on the dataset, with the goal of predicting whether the articles are truthful or fake. The experiments started by studying the impact of the size (n) of n-grams on the performance. We started with unigram (n = 1), then bigram (n = 2), then steadily increased n by one until reaching n = 4. Furthermore, each n value was tested combined with a different number of features. The experiments were run using 5-fold cross validation; in each validation round the dataset is divided into 80% for training and 20% for testing. The algorithms were used to create learning models, and then the learned models were used to predict the labels assigned to the testing data.

4.3 EXPERIMENT RESULTS

We studied two different features extraction methods, TF-IDF and Count Vectorizer, and varied the size of the n-gram from n = 1 to n = 4. We also varied the number of features p (i.e., top features selected), ranging from 1,000 to 5,000. Tables 1, 2, 3, 4, 5 and 6 show the obtained results.

Table1.MultinomialNB

N GRAM SIZE	TF-IDF		COUNT VECTORIZER	
	1000	5000	1000	5000
Uni Gram	0.850	0.867	0.875	0.897
Bi Gram	0.665	0.746	0.665	0.745
Tri Gram	0.571	0.652	0.574	0.652
Four Gram	0.552	0.594	0.596	0.553

Table 2. Gradient Boosting

N GRAM SIZE	TF-IDF		COUNT VECTORIZER	
	1000	5000	1000	5000
Uni Gram	0.917	0.917	0.918	0.918
Bi Gram	0.813	0.813	0.813	0.815
Tri Gram	0.796	0.795	0.796	0.871
Four Gram	0.557	0.557	0.557	0.867

Table 3. Random Forest Classifier

N GRAM SIZE	TF-IDF		COUNT VECTORIZER	
	1000	5000	1000	5000
Uni Gram	0.926	0.935	0.925	0.934
Bi Gram	0.854	0.872	0.855	0.871
Tri Gram	0.803	0.816	0.803	0.816
Four Gram	0.644	0.691	0.664	0.691

Table 4. Passive Aggressive Classifier

N GRAM SIZE	TF-IDF		COUNT VECTORIZER	
	1000	5000	1000	5000
Uni Gram	0.906	0.913	0.900	0.916
Bi Gram	0.850	0.872	0.851	0.867
Tri Gram	0.804	0.818	0.803	0.815
Four Gram	0.554	0.596	0.554	0.592

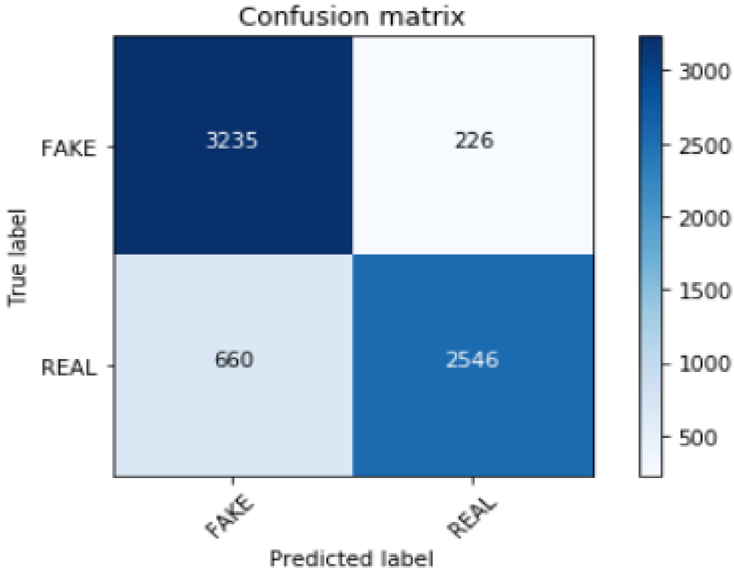
Table 5. Multinomial Classifier with Hyperparameter (alpha 1)

N GRAM SIZE	TF-IDF		COUNT VECTORIZER	
	1000	5000	1000	5000
Uni Gram	0.850	0.886	0.874	0.897
Bi Gram	0.666	0.746	0.668	0.747
Tri Gram	0.572	0.651	0.575	0.652
Four Gram	0.522	0.594	0.552	0.596

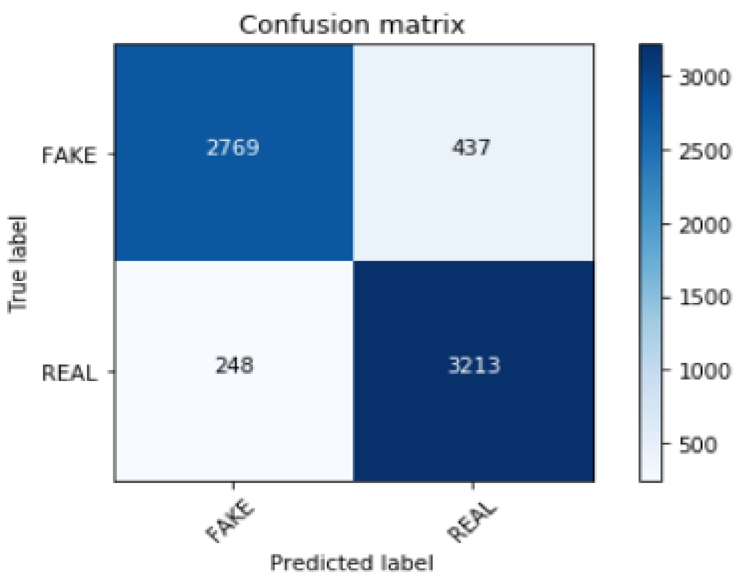
Confusion Matrices

For Multinomial NB -

With count Vectorizer:

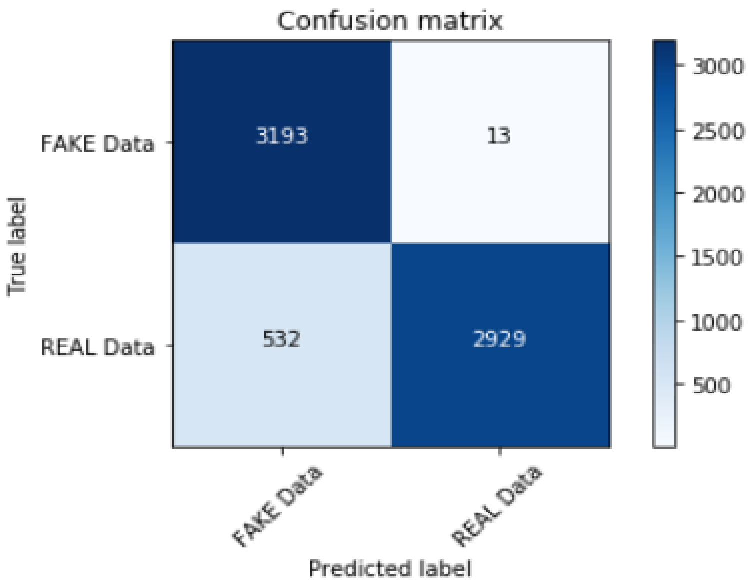


With TfidfVectorizer:

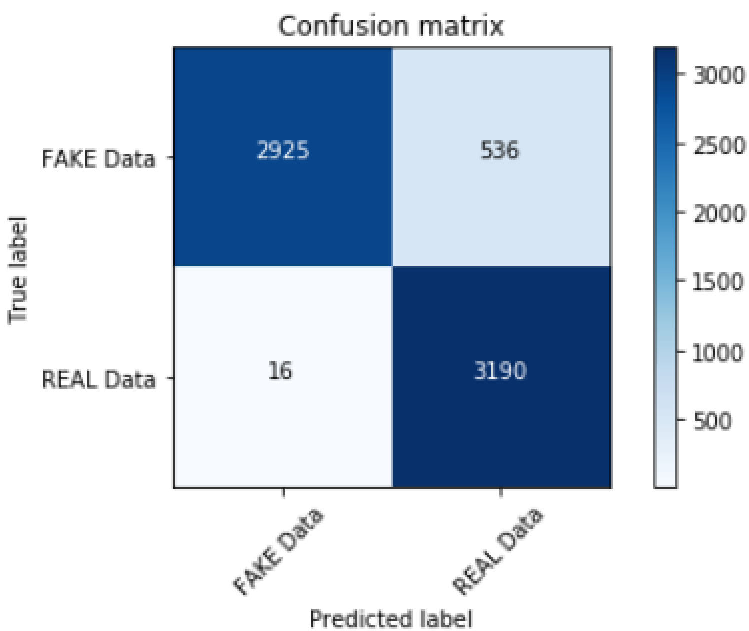


For Gradient Boosting -

With count Vectorizer:

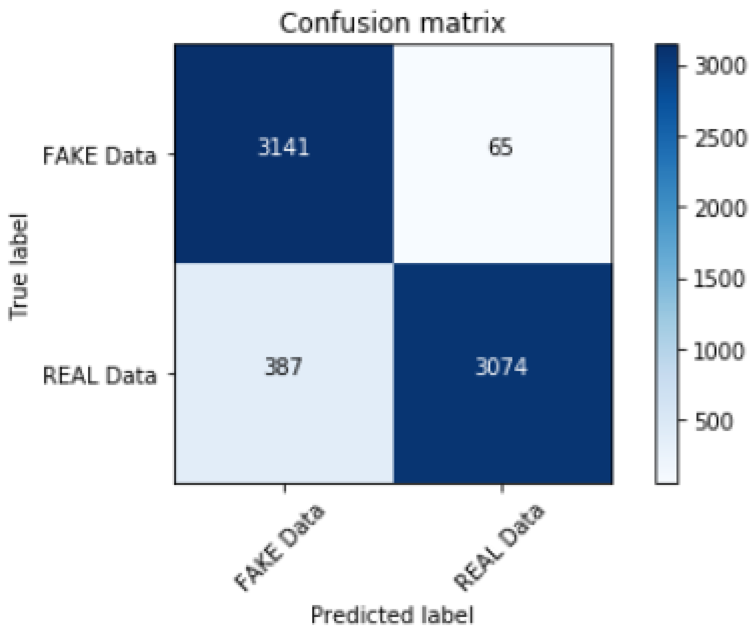


With tfidfVectorizer:

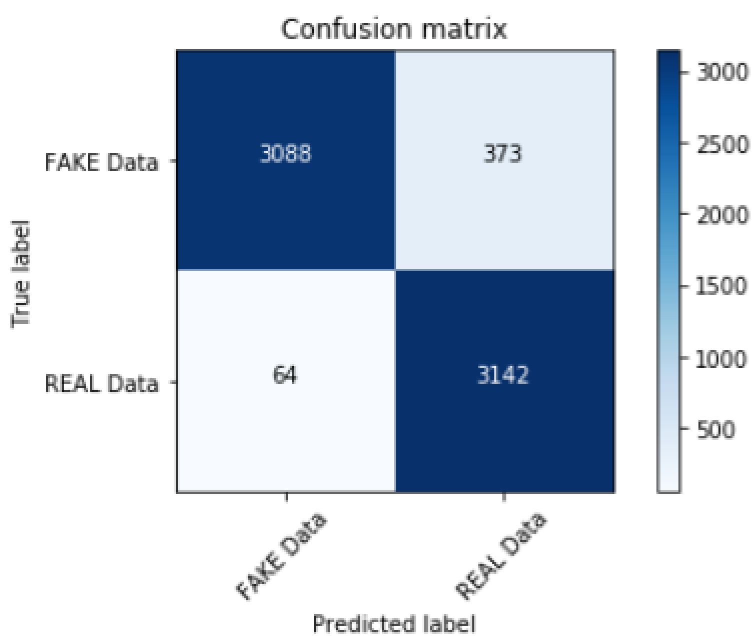


For Random Forest Classifier -

With countVectorizer:

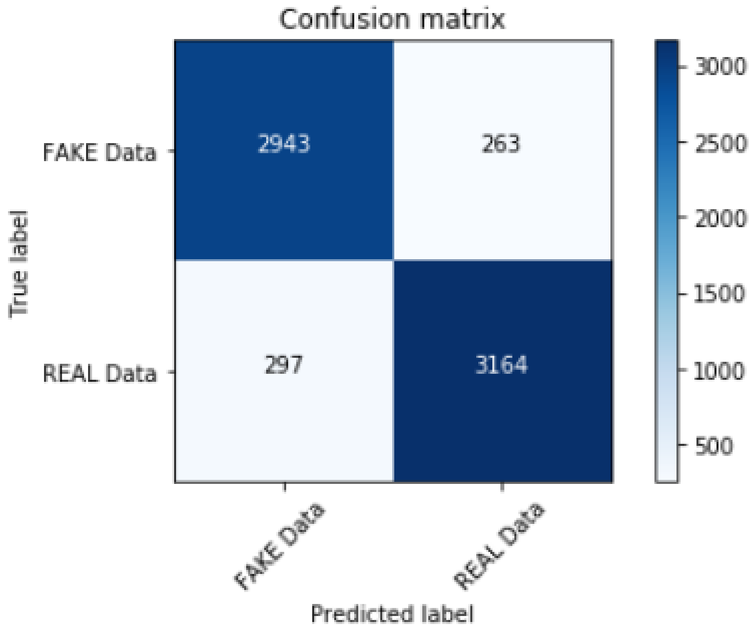


With tfidfVectorizer:

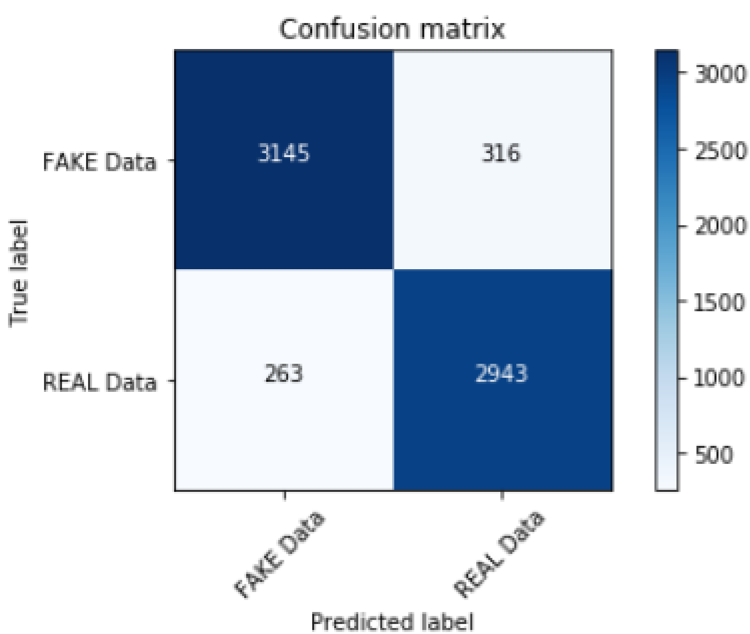


For Passive Aggressive Classifier -

With countVectorizer:



With TfidfVectorizer:



5. CONCLUSION

The problem of fake news has gained attention in 2016, especially in the aftermath of the last US presidential elections. Recent statistics and research show that 62% of US adults get news on social media [12, 13]. Most of the popular fake news stories were more widely shared on Facebook than the most popular mainstream news stories [14]. A sizable number of people who read fake news stories have reported that they believe them more than news from mainstream media. Dewey [15] claimed that fake news played a huge role in the 2016 US election and that they continue to affect people opinions and decisions.

In this paper, we have presented a detection model for fake news using n-gram analysis through the lenses of different features extraction techniques. Furthermore, we investigated two different features extraction techniques and five different machine learning techniques.

The proposed model achieves its highest accuracy when using unigram feature and Random Forest classifier. The highest accuracy score is 93.5% for 5000 feature values. This classifier performs well no matter the number of feature values used. Also, with the increase of n-gram (Tri-gram, Four-gram), the accuracy of the algorithm decreases. The lowest accuracy of 52.2% was achieved using multinomial classifier with hyperparameter(alpha=0.1) with four-gram words for 1,000 feature values.

REFERENCES

1. Schow, A.:The 4 Types of 'Fake News'. Observer (2017). http://observer.com/2017/01/fake-news-russia-hacking-clinton-loss/

2. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST 2015). Article 83, p. 4, American Society for Information Science, Silver Springs (2015)

3. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: the 2nd International Workshop on News and Public Opinion at ICWSM (2017)

4. Bur foot, C., Baldwin, T.: Automatic satire detection: are you having a laugh? In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 4 August 2009, Suntec, Singapore (2009)

5. Wang, W.Y.: Liar, Liar Pants on fire: a new Benchmark dataset for fake news detection. arid preprint (2017). arXiv:1705.00648

6. Rubin., Victoria, L., et al.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of NAACL-HLT (2016)

7. The Verge: Your short attention span could help fake news spread (2017). https://www.theverge.com/2017/6/26/15875488/fake-news-viral-hoaxes-botsinformation-overloadtwitter-facebook-social-media. Accessed 16 Aug 2017

8. Lemann, N.: Solving the Problem of Fake News. The New Yorker (2017). http://www.newyorker.com/news/news-desk/solving-the-problem-of-fake-news
Detection of Online Fake News Using N-gram Analysis 137

9. Schulten, K.: Skills and Strategies—Fake News vs. Real News: Determining the Reliability of Sources. The Learning Network (2017). https://learning.blogs.nytimes.com/2015/10/02/skills-and-strategies-fake-news-vs-real-news-determining-the-reliability-of-sources/. Accessed 16 Aug 2017

10. Levin, S.: Pay to sway: report reveals how easy it is to manipulate elections with fake news. The Guardian (2017). https://www.theguardian.com/media/2017/jun/13/fake-news-manipulate-elections-paid-propaganda

11. Gu, L., Kropotov, V., Yarochkin, F.: The fake news machine, how propagandists abuse the internet and manipulate the public. In: 1st ed. [pdf] Trend Micro, p. 81 (2017). https://documents.trendmicro.com/assets/white_papers/wp-fake-news-machine-how-propagandists-abuse-the-internet.pdf?_ga=2.117063430.1073547711.1497355570 1028938869.1495462143

12. Gottfried, J., Shearer, E.: News use across social media platforms. Pew Res. Cent. 26 (2016)

13. Gottfried, J., et al.: The 2016 presidential campaign–a news event that's hard to miss. Pew Res. Cent. 4 (2016)

14. Silverman, C., Singer-Vine, J.: Most americans who see fake news believe it, new survey says. BuzzFeed News (2016)

15. Dewey, C.: Facebook has repeatedly trended fake news since firing its human editors. Washington Post (2016)