Fall 2021

A Comparative Study on Feature Extraction and Classification/ Clustering of Fake News and Conspiracy Theories from Twitter Data

Deb Shana
Shana_Deb@coumbusstate.edu

COLUMBUS STATE UNIVERSITY

A COMPARATIVE STUDY ON FEATURE EXTRACTION AND CLASSIFICATION/CLUSTERING OF FAKE NEWS AND CONSPIRACY THEORIES FROM TWITTER DATA

A THESIS SUBMITTED TO THE TURNER COLLEGE OF BUSINESS

IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED COMPUTER SCIENCE

TSYS SCHOOL OF COMPUTER SCIENCE

BY

SAHANA DEB

COLUMBUS, GEORGIA

2021

A COMPARATIVE STUDY ON FEATURE EXTRACTION AND CLASSIFICATION/CLUSTERING OF FAKE NEWS AND CONSPIRACY THEORIES FROM TWITTER DATA

By

Sahana Deb

Committee Chair:

Dr. Lydia Ray

Committee Members:

Dr. Rania Hodhod

Dr. Lixin Wang

Columbus State University
October 2021

ABSTRACT

Fake news and conspiracy theories have become largely abundant in the expanding world of social media. They predominantly affect the beliefs and thoughts of the public, resulting in chaos. They have always existed throughout the last few decades. They have been linked to prejudice, revolutions and genocide across history. They have also been known to have propelled people to reject mainstream medicines to an extent where some diseases are recurring in some parts of the world. They impose a serious impact since they are capable of spreading very fast Thus, it is very important to find suitable ways to detect fake news and conspiracy theories in social media, which requires a thorough analysis of their features. This study presents a survey on the various techniques of feature extraction and classification that can be implemented to classify and detect fake news and conspiracy theories from twitter datasets. The results indicate that the tf-idf method of feature extraction, when implemented with the svm classification algorithm, yields the highest accuracy of 99.6% in comparison to the other algorithms i.e. multinomial naive bayes, logistic regression and decision tree. The Bag of Words model yields an accuracy of 52.3% for both multinomial naive bayes and logistic regression algorithms and a lower range of accuracies for the other two algorithms i.e. svm and decision tree . TF-IDF has thus performed better than Bag of Words.

Keywords: conspiracy theories, fake news, feature extraction, Machine Learning, algorithms

ACKNOWLEDGEMENTS

I would like to thank my professors, Dr. Lydia Ray and Dr. Rania Hodhod for being wonderful guides and mentors. Their immense support has proved to be one of the greatest motivations for my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

A COMPARATIVE STUDY ON FEATURE EXTRACTION AND CLASSIFICATION/CLUSTERING OF FAKE NEWS AND CONSPIRACY THEORIES FROM TWITTER DATA

A thesis submitted to Turner College of Business in partial fulfillment of the requirements for the degree of

MASTER OF APPLIED COMPUTER SCIENCE

TSYS SCHOOL OF COMPUTER SCIENCE

By

Sahana Deb

2021

_____	_____
Dr. Lydia Ray, Chair	Date
_____	_____
Dr. Rania Hodhod, Member	Date
_____	_____
Dr. Lixin Wang, Member	Date

Chapter I. Introduction

1. Introduction

Social media has a tremendous impact on the thoughts and beliefs of the public. Besides being a good source of information, it also comprises news and stories which are based on rumors and conspiracies. People tend to turn to social media as informational sources from time to time. Generally, rumors on social media tend to emerge during situations of crisis or during the occurrence of major events, such as the outbreak of a global pandemic, presidential elections, war, etc. [1]. Similar situations also generate conspiracy theories, which can further confuse people, rather than helping them understand the situation with correct information. According to the authors of "Understanding Conspiracy Theories" [2], "conspiracies typically attempt to usurp political or economic power, violate rights, infringe upon established agreements, withhold vital secrets, or alter bedrock institutions." The conspiracy theories are created in an attempt to explain the eventual causes of significant societal or political events by claiming the existence of secret plots and ideas [3]. These theories tend to make difficult situations easily understandable through creating suspicions that powerful people and organizations are misleading or tricking the public, by means of their evil plans [4]. Hence, it can be stated that rumors and conspiracy theories easily attract the public at large, especially when spread through social media.

Conspiracy theories are identifiable by human beings, though this process takes time and might also be confusing at times. The influence of the Internet and social media has created a serious issue out of conspiracy theories and fake news, mainly because of two reasons. Firstly, fake

news and conspiracy theories spread very fast and reach millions of people within seconds. Also, social media platforms allow free speech with almost no censorship. Therefore, fake news/conspiracy theory is easily published and shared at a magnitude which is impossible to be controlled manually. For example, shortly after the pandemic started, a false claim emerged and spread through social media that the coronavirus started in a lab in Wuhan, China, since the pathogen first emerged there. Whilst this claim has been denied by U.S. intelligence agencies, a large number of people believed it as it spread through social media [5]. Another popular conspiracy theory that emerged during the pandemic was the unproven claim that National Institute of Allergy and Infectious Diseases director Anthony Fauci and Microsoft co-founder Bill Gates could be using their power to profit from a COVID-19 vaccine [6]. It was also claimed that the virus emerged in a lab and that wearing masks could increase the chances of contracting it. These unwarranted claims were made by Judy Mikovits, a former researcher, who was featured in the conspiracy theory film, “Plandemic”. An excerpt from the film was shared by the conspiracy theory group Qanon, and the video was viewed on social media, more than eight million times [7]. These examples demonstrate that the Internet, and most specifically social media can amplify conspiracy theories/fake news at an unprecedented scale beyond any human control.

1.1 Problem Statement

It is evidently important to propose computational methods that can differentiate fake news and conspiracy theories from real facts. These methods would be helpful to support fact checking organizations and help recognize and prevent the spread of misleading information among the public [8]. Data driven artificial intelligence shows promise in classification of large amounts of unstructured data. Machine Learning based classification algorithms need to be explored to

develop a solution to the problem of automatic classification of fake news/conspiracy theories. Furthermore, conspiracy theories/fake news are characterized by certain linguistic features that can differentiate these from articles presenting facts. Therefore, Natural Language Processing (NLP) methods can be used in the automatic detection of conspiracy theories and fake news as NLP plays an important role in the analysis of linguistic features. Hence, employing various NLP techniques combined with Machine Learning algorithms can lead to more accurate classification of fake news/conspiracy theories. In this research project, a comparative study on the various feature extraction is conducted and classification techniques of conspiracy theories are proposed. Machine Learning is an extremely powerful tool to make predictions. Supervised learning is useful in the classification of large labelled datasets. A large twitter dataset, containing tweets about the Covid 19 pandemic will be used for the study.

1.2 Features of Conspiracy Theories

Conspiracy theories are attempts to explain the ultimate causes of significant social and political events and circumstances with claims of secret plots by two or more powerful actors [9]. There are certain linguistic features in conspiracy theories that make them different from real facts, some of which are mentioned below:

- Involvement of a hypothesized pattern, demanding that the plans of the alleged conspirators are intentional. For example, a conspiracy theory which claimed that major medical professionals and pharmacies such as Big Pharma have already found the cure for cancer and are withholding it [10].
- Containing an element of threat like the goals of the conspirators are harmful and deceptive. For example, some supporters of alternative medicines believe that major drug

companies that dominate the drug industry conspire to keep people sick in order to gain profits by hiding important medical information from the public [11].

- Carries an element of secrecy, which makes them difficult to invalidate. For example, in 2001, a program called, "Conspiracy Theory: Did We Land on the Moon?" was aired on Fox Television, which questioned the authenticity of the Apollo moon landing in 1969, by rehashing several inconsistencies between the official version of the moon landing and its photographs [12].

Several conspiracy theories can be found while browsing through the Internet [13]. They mostly involve powerful and influential groups such as groups of politicians or influential business people, all of whom are thought to be conspiring towards evil goals [14]. The numerous conspiracy theories regarding Princess Diana's death in 1997 were so convincing and widespread, that the Met Police was forced to launch an inquiry called "Operation Paget" in order to find if there were any truths in the theories. Almost 175 theories were examined, many of which were supported by the "Daily Express" [15].

Another important feature of conspiracy theories is that they spread rapidly through social media, which also makes them extremely dangerous. For example, a theory suggested the malaria drug, hydroxychloroquine, was an effective treatment for the coronavirus. It was strongly supported by former US President, Donald Trump, which made it more believable to the public.

According to the European Commission and UNESCO [16], most conspiracy theories have several things in common, such as an alleged script or secret plot, a group of conspirators, and evidence that seems to support the theory and false suggestions that nothing happens by accident and that there are no coincidences. It is highly unlikely that the author of [16] used

verifiable facts and evidence from scientific research and academic records. In most cases, the source of the information is not clear and the tone is subjective and emotionally charged.

One of the most bizarre conspiracy theories, QAnon, originated in the USA, has quickly spread in Europe during the pandemic. It has been found that the pandemic has acted as a catalyst in boosting its popularity across Europe. Several QAnon placards were featured across Europe during protests against coronavirus restrictions, mainly in Berlin, London and Paris.

1.3 Thesis Goals

The aim of this research is to study various feature extraction techniques and how effective they are when implemented in combination with various classification algorithms on a dataset. A comparative study has been conducted based on the performances of the various techniques along with the algorithms, on a large twitter dataset. The two feature extraction techniques used in this research are the bag of words and tf-idf vectorizer. The four selected classification algorithms include multinomial Naive Bayes, support vector machine (SVM), logistic regression and decision tree. The corresponding accuracies and confusion matrices for the respective implementations have been recorded.

1.3 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 provides a survey on existing works and methodologies used; Chapter 3 consists of the methodologies including the proposed approach, dataset description, dataset processing, implemented techniques, and results and analysis of results. Finally, Chapter 4 provides the discussion, conclusion, and future work.

Chapter II. Survey on Existing Works

2.1 Literature Review

Several studies have been published in this field. Detection of fake news and conspiracy theories has emerged to be a necessity.

M. Wood in [17] investigated the characteristics of conspiracy theories that originated during the Zika outbreak of 2015-2016 on Twitter. An adaptive version of the Rumor Interaction Analysis System (RIAS) has been implemented, which allows quantitative classification of rumor-spreading messages. The messages are examined for the expression of belief or disbelief in a rumor, whether it contains a directive or shows an attempt to authenticate the information in it. The sample included 25.162 original tweets that referred to at least one Zika conspiracy, among which, 17,421 expressed belief, 6,555 expressed disbelief and 1186 were ambivalent. The analysis performed by the adaptive RIAS disclosed significant differences between belief and disbelief tweets in terms of authentication (belief - 25.56% and disbelief - 5.80%) and rhetorical questions (belief - 14.90% and disbelief - 9.37%).

E. Ferrara in [18] investigated the evidence of the presence of automated bots in twitter, in the online discussion about the Covid-19 pandemic. This article also studies the prevalence, behavioral characteristics and volume of activity of the bots compared to that of the human accounts, if the evidence of their presence is found. It analyses the role of bots in pushing ideologies and political narratives in social media. Based on prior research, where the role of bots in pushing ideologies and political narratives in social media has been demonstrated, the authors pose a second research question about observing any pattern of preferential behavior where the bots seem to focus on fueling specific topics of discussion concerned with politics or ideology. Furthermore, bot score analysis has been performed to report six basic account meta-data features that can help predict the differentiation between bots and human users. Then, the

authors perform an age and provenance analysis on the accounts according to their bot scores. In order to address the second research question, they have used two distinct strategies, namely, keywords and hashtag analysis. Furthermore, the authors mention that the detection of bots is a difficult task and even refined Machine Learning algorithms produce fluctuating levels of accuracy.

A-All Tanvir et al in [19] proposed a model for the detection of fake news from twitter posts and have performed a comparison between five well known Machine Learning algorithms, namely, Support Vector Machine, Naïve Bayes Method, Logistic Regression, Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) and demonstrated the classification efficiency for each one of them. The results imply that the Naive Bayes Model, when tested on the defined feature vectors, produces accuracies of 73% on count vector feature, 75% or Word Level TF-IDF, N-gram vector and character vectors. The Logistic Regression Model produced accuracies of 74% and 76% on count and word level vectors respectively. Next, the SVM Model produces an accuracy of 74% in all four feature vectors. In case of deep learning methods, accuracies of 73% and 74% were achieved for LSTM and RNN respectively.

Karen M. Douglas et al in [20], discussed how conspiracy theories spread through social media platforms and are shared among people.

The authors in [21] applied a narrative framework discovery pipeline to several social media posts and news stories in order to detect conspiracy theories. The data for this study has been derived by concatenating several social media resources and Covid-19 related news stories from reputable journalistic resources.

The authors of [22] presented an NLP based technique for the detection of Covid-19 misinformation videos on YouTube by the analysis of user comments. They have performed

multi-label classification in order to classify the content and have used classification models such as logistic regression, SVM and random forest.

2.2 Methodologies Used in Existing Research

Feature engineering plays an essential role in the analysis of conspiracy theories. Extracting and adjusting the features of a large dataset would ensure a good classification accuracy. It is also one of the best ways to reduce dimensionality of a large dataset. Feature extraction methods are used either separately or simultaneously in order to improve performances such as accuracy, visualization and readability of acquired knowledge [23]. Through the feature extraction methods, some original features of the dataset are transformed into more significant other features [24]. Some of the NLP methods and Machine Learning algorithms that have been implemented in the above-mentioned research studies are discussed below.

Methods of processing the datasets: In order to apply the classification algorithms, it is first necessary to process the datasets and extract important features, which can then be fed to the algorithms. Some of the NLP techniques to be used for feature extraction and processing the dataset are as follows:

1. Bag of Words (BoW): It is one of the most fundamental NLP techniques where the data is transformed into tokens which are further transformed into a set of features. The BoW model is useful for data classification, where each word is treated as a feature. Initially, the dataset is converted into lowercase and then all punctuations and unnecessary symbols are removed. A vocabulary of words is then formed, which is used to create a dictionary, which includes the frequency of each word as they appear in the document. A set of unique words are extracted from the dictionary. Next, a matrix of features is

formed where each word is allocated a column and values are assigned to them according to their occurrence in the data. This process is called text vectorization.

The BoW model can be implemented on a Covid-19 dataset, which would produce a matrix consisting of frequent words from the dataset. The matrix can then be fed to a classifier for further analysis.

2. TF-IDF Vectorizer: Term frequency - inverse document frequency states the frequency of a term according to its occurrence in the entire dataset [25]. A metric value is assigned to represent that term [26]. This value also says how important the term is.

   The term frequency can be calculated as follows [27]:

   TF = number of times a term appears in the dataset or total number of the terms in the dataset. Depending on the different input types, various TF-IDF scores can be generated such as word-level TF-IDF, N-gram level TF-IDF, character level TF-IDF [28].

   The TF-IDF scores generated from a Covid-19 dataset could help achieve a clearer picture of the relevant words that have been used in the conspiracy theories. This would further enhance the likelihood of obtaining more accurate results after the linguistic analysis.

3. word embedding: This method preserves the context and relationships of words in the dataset, through the vector space model. This makes detection of similar words easier. There are various implementations of word embedding such as word2vec, GloVe, FastText, etc.

4. Principal Component Analysis: PCA is one of the most popular and widely used feature extraction methods [29]. It is a simple non-parametric method that is implemented in order to extract useful features from large redundant datasets [30].

Methods of Classification: Machine Learning provides us with several classification algorithms.

1. Naïve Bayes classification: This is based on Bayes' Theorem which determines the probability of a hypothesis, based on its prior probability. This classification algorithm requires a part of data to be trained and the other part is then tested based on the parameters obtained from the training. It is pretty fast.
2. Logistic Regression: This classification algorithm implements a logistic function to model the probabilities representing the possible outcomes of an event.
3. Support Vector Machine (SVM): It is a supervised Machine Learning algorithm where each data item is plotted in n-dimensional space as a particular point and the value of each feature is considered to be the value of a particular coordinate. The hyperplane that differentiates the two classes is determined to perform the classification.
4. Decision Tree: This learning algorithm compiles training data in a tree-like structure. Each branch of the flowchart represents the relationship between feature values and the class label [31]. The decision tree learns from a set of training data in an iterative process [32]. Entropy measure for each feature is calculated and the probabilities are estimated in a similar manner as Naive Bayes.

Chapter III. Methodology

3.1. Proposed approaches

This research proposes a comparative analysis on the different feature extraction methods implemented on the various classification algorithms. The feature extraction methods that

have been implemented are the bag of words model (BoW) and the term frequency-inverse document frequency model (TF-IDF). The classification algorithms used in this research include

- Multinomial Naive Bayes
- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)

3.2 Dataset Description

A dataset consisting of 44,921 tweets has been collected. Initially, the tweets were categorized into two folders: true and fake, respectively. There are 21,418 tweets in the ‘true news’ category and 23,503 tweets in the ‘fake news’ category. The csv files contain information like ‘title’, ‘text’, ‘subject’ and ‘date’ which are presented as column headers.

3.3 Dataset Processing

The dataset is processed and a new csv file is created to store the processing outputs. The tweets from both folders (true and fake) were integrated into one csv file. The columns, ‘title’ and ‘subject’ were extracted and retained in the modified csv file. A new column, indicating the tweet category (true or fake) was added. The following figures depict the layout of the dataset:

	A	B	C	D	E
1	title	text	subject	date	Category
2	As U.S. buc	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, wh	politicsNev	31-Dec-17	T
3	U.S. militai	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S	politicsNev	29-Dec-17	T
4	Senior U.S.	WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President	politicsNev	31-Dec-17	T
5	FBI Russia	WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplo	politicsNev	30-Dec-17	T
6	Trump wai	SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on F	politicsNev	29-Dec-17	T
7	White Hou	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - The White House said on Friday it was set to k	politicsNev	29-Dec-17	T
8	Trump say	WEST PALM BEACH, Fla (Reuters) - President Donald Trump said on Thursday he believes he will b	politicsNev	29-Dec-17	T
9	Factbox: T	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald	politicsNev	29-Dec-17	T
10	Trump on	The following statementsÂ were posted to the verified Twitter accounts of U.S. President Donald	politicsNev	29-Dec-17	T
11	Alabama o	WASHINGTON (Reuters) - Alabama Secretary of State John Merrill said he will certify Democratic	politicsNev	28-Dec-17	T

Figure 3.1 : Dataset Structure

23147	Trump HU	Is The Donald really this oblivious to the entire world? And history? And, well, everything? Two an	News	20-Apr-17	F
23148	WATCH: T	Canadian Prime Minister Justin Trudeau just demonstrated that he s the adult in the room when i	News	20-Apr-17	F
23149	GOP FURI	Not too long ago, the Republican Party and Donald Trump made a disgusting move to screw over	News	20-Apr-17	F
23150	Chaffetz S	House Oversight Committee Chairman Jason Chaffetz (R-Utah), doesn t think he wants to be a Co	News	20-Apr-17	F
23151	AG Sessior	Trump s Attorney General Jeff Sessions is as furious as Trump is over the blocking of an executive c	News	20-Apr-17	F
23152	Sarah Pali	Sarah Palin, Kid Rock, and Ted Nugent visited the White House on Wednesday. No, that s not a se	News	20-Apr-17	F
23153	Voters Pre	The current occupant of the White House has finally accomplished something for his first 100 days	News	20-Apr-17	F
23154	Exxon Mol	During Watergate, the line became follow the money. In the Trump, Russia scandal, you can bet	News	20-Apr-17	F
23155	New Rese	Build the wall! Build the wall! Supposedly nobody builds better walls than Donald Trump, and the	News	19-Apr-17	F
23156	Patriots W	It s no secret that Donald Trump has deep insecurities about his job performance compared to his	News	19-Apr-17	F
23157	BREAKING	While we don t yet have the smoking gun that specifically says Donald Trump colluded with the Ru	News	19-Apr-17	F

Figure 3.2: Processed Dataset

Figure 1 reflects the overall structure of the dataset displaying the various columns: title, text, subject, date and category respectively. Figure 2 represents the dataset after it is processed. The unnecessary columns have been removed and only the columns representing the main data (tweet text and tweet category) are kept. This processed dataset is further used for feature extraction and then fed to the classification algorithms.

The ‘text’ column was extracted from the csv file and all special characters and single characters were removed from the text. Also, the multiple spaces have been substituted by single spaces and all text has been converted into lower case.

3.4 Applied Techniques

Once the dataset is processed by extracting the text categorically and then processing it in the above-mentioned ways, the Bag of Words model is implemented on the text.

Initially, a bag-of-words representation was one of the most popular representation methods for object categorization. As described by the authors of [33], the main idea is to quantize each extracted key point into one of the visual words, and then to represent each image as a histogram of the visual words. Normally, a clustering algorithm is implemented to generate the words.

In this study, the count-vectorizer tool has been used to implement the bag of words model. This tool, available in the scikit-learn library in Python, is used to transform a given text into vectors on the basis of the word frequencies, as per their appearances throughout the entire text. It creates a matrix where each column is represented by each unique word and each document represents each row in the matrix. The “max_features” parameter of the count-vectorizer indicates the number of features, ordered by term frequency, that are to be considered in forming the vocabulary.

Furthermore, the dataset is split in a ratio of 7:3 for training the classifier. 70% of the data is used for training and the classifier is tested on the remaining 30% of the data. Along with the accuracy, other metrics like f1 score, precision score, recall score and confusion matrix have been recorded for each of the classification algorithms.

Next, another approach is adopted by implementing the TF-IDF (term frequency - inverse document frequency) model to the preprocessed data before it is fed to the classifier for training and testing purposes. This model converts a collection of raw documents into a matrix of TF-IDF features. The count vectorizer tool of python in the scikit-learn library is also implemented with this model to extract the features and form the count matrix which has been further fed to the TF-

IDF transformer to convert the matrix into normalized TF-IDF representation. The matrix is then used to train and test the different classifiers as before. The same metrics are evaluated as before.

3.5 Results

The results obtained for the first model implementing bag of words are as below:

Table 3.1: Results for bag of words model

Algorithm	**Accuracy**	**F1 score**	**Precision score**	**Recall score**	**Confusion matrix**
Multinomial NB	0.523	0.442	0.512	0.506	[[6088 972] [5458 958]]
SVM	0.509	0.498	0.503	0.503	[[4439 2621] [3993 2423]]
Logistic Regression	0.523	0.417	0.510	0.504	[[6393 667] [5764 652]]
Decision Tree	0.503	0.503	0.504	0.504	[[3408 3652] [3049 3367]]

The results obtained for the second model implementing TF-IDF are as below:

Table 3.2: Results for TF-IDF model

Algorithm	**Accuracy**	**F1 score**	**Precision score**	**Recall score**	**Confusion matrix**

Multinomial NB	0.945	0.945	0.945	0.944	[[22415 1086] [1398 20020]]
SVM	0.996	0.996	0.996	0.996	[[7028 32] [22 6394]]
Logistic Regression	0.514	0.504	0.509	0.509	[[4408 2643] [3906 2519]]
Decision Tree	0.504	0.501	0.501	0.501	[[3873 3178] [3511 2914]]

3.6 Analysis of Results

The first model, in which BoW was implemented on the data prior to being fed to the classifiers, produced low accuracies in the range of (50-52) %. The second model, in which TF-IDF has been implemented on the data prior to being fed to the classifiers, produces high accuracies for the multinomial naive bayes algorithm (94%) and SVM (99.6%) respectively, and a series of low accuracies for the other two algorithms. The following graphs represent the accuracies for all the models respectively:

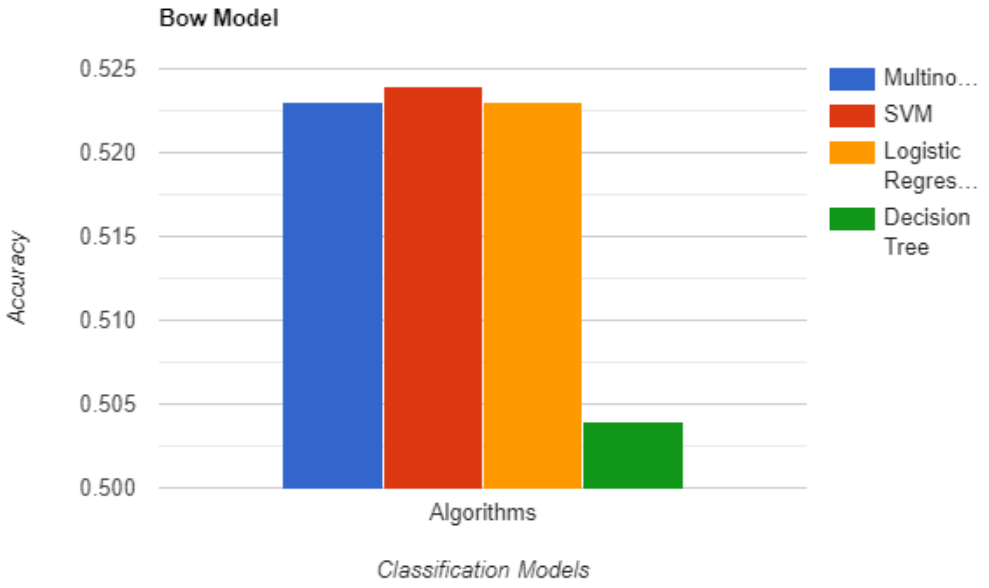

Figure 3.3: Graph representing accuracies for algorithms implementing the BoW model

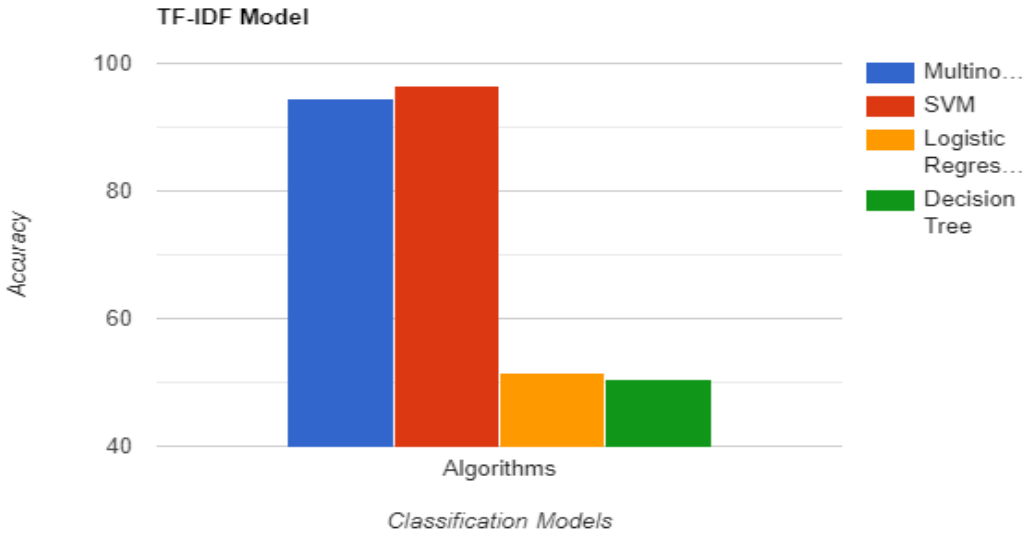

Figure 3.4: Graph representing accuracies for algorithms implementing the TF-IDF model

The bag-of-words model, albeit simple to generate, is far from perfect. The model tends to ignore the positional statistics of the words throughout the text which affects the accuracy as the

location information of words poses to be an important piece of information. [34] Similar words, used to express different sentiments are allotted the same vectorial representation. Also, this model seldom takes into account the semantics of the words in the corpus, which in the allotment of totally different vectorial representations to similar words, used in similar contexts. All these factors mostly lead to low accuracy.

On the other hand, the TF-IDF model accumulates more information on the important as well as the less important words, unlike the BoW model. It is known to perform better with the classification models since it takes into account a normalized frequency of the words instead of the raw count. Also, the TF-IDF can eliminate uninformative words unlike the BoW since it can both “stretch” and “compress” the word count making some of them higher and some lower as required [35].

In order to evaluate the classification performance of algorithms, the metrics, Precision and Recall, also referred to as evaluation metrics are often used. These metrics are calculated using the confusion matrix.

The precision in a classification problem is referred to as the ability of the classification model to identify only the relevant data points [36]. It is defined as the ratio of the number of true positives over the sum of the number of true positives and the number of false positives. Recall is referred to as the ability of the model to find all the relevant cases within the corpus [37]. It is defined as the ratio of the number of true positives over the sum of the number of true positives and the number of false negatives. It is through recall that the model finds all the relevant data points within the dataset. In our study, the multinomial svm classifier model with TF-IDF for feature extraction, has precision and recall scores as 0.996 and 0.996, respectively, both of which

are significantly higher compared to the rest of the models. According to the confusion matrix of the same, the following features have been noted:

True Positive : 7028

False Positive : 32

False Negative : 22

True Negative : 6394

We can further derive from the confusion matrix that this model yields a high sensitivity and specificity of 0.9969 and 0.995, respectively. A high sensitivity denotes the stronger ability of the model to predict the true positives in every category. A high specificity denotes that the number of false positives are low. This explains the better results delivered by the TF-IDF model when implemented with the svm classifier. Also, the multinomial naive bayes classifier has yielded a good accuracy of 0.945, when implemented alongside the TF-IDF model. The recorded sensitivity and specificity were, 0.941 and 0.949, respectively.

Chapter IV. Discussion and Conclusion

4.1. Discussion

The results achieved in this study indicate that the TF-IDF model for feature extraction, when implemented along with the svm classifier, yields the highest accuracy (99.6%) followed by the multinomial naive bayes classifier (94.5%). The other two classification algorithms, i.e. logistic regression and decision tree, yield much lower accuracies, when paired with the TF-IDF model. The bag of words model, when implemented with all four classifiers, yields consistently low accuracies (between 50-52%). The TF-IDF method of feature extraction has thus shown better results, when applied to a large dataset, in comparison to the bag of words method.

Amidst all four classification algorithms, the svm classifier has provided the highest accuracy. It creates a hyperplane in an N-dimensional space (where N is the number of features). This hyperplane or line separates the data into classes and svm uses the kernel in order to find the best line hyperplane that separates the classes; which lowers the risk of error on the data. [38] The large margin that is generated, allows the fitting of more data and their classification perfectly. [39] Followed by the svm classifier, the multinomial naive bayes classifier has provided a high frequency too. The facts that it can be only used for textual data classification and that it is highly scalable along with the ability of handling large datasets (similar to the large dataset used in this study), aid to its better results.

4.2. Conclusion

It can be concluded that the TF-IDF method of feature extraction has yielded much higher accuracies, when applied to certain classification algorithms (svm and multinomial naive bayes), in comparison to the bag of words method. The former's better performance can be accredited by the fact that it contains more information on words of both high and low importance respectively, in comparison to the latter.

4.3 Future Work

This work paves the way for a lot of future research in this discipline. The need for feature extraction techniques is ever growing in the field of natural language processing as they play a crucial role in the learning procedure of the algorithms from a predefined set of features, in order to produce output for the test data. The accuracy of the results also depends heavily on the feature extraction methods. The two methods used in this study, Bag of words and TF-IDF, are

among the most popular methods of feature extraction. However, there are several other methods that could possibly yield good results. For example, word embedding algorithms like GloVe and Word2Vec, Principal Component Analysis (PCA) and one-hot encoding.

References

1. Wood, M. J. (2018). Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. *Cyberpsychology, behavior, and social networking*, *21*(8), 485-490.

2. Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3-35.

3. Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3-35.

4. Wood, M. J. (2018). Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. *Cyberpsychology, behavior, and social networking*, *21*(8), 485-490.

5. Lewis. T. (2020, October 12). Eight Persistent COVID-19 Myths and Why People Believe Them. Scientific American. https://www.scientificamerican.com/article/eight-persistent-covid-19-myths-and-why-people-believe-them/

6. Lewis. T. (2020, October 12). Eight Persistent COVID-19 Myths and Why People Believe Them. Scientific American. https://www.scientificamerican.com/article/eight-persistent-covid-19-myths-and-why-people-believe-them/

7. Lewis. T. (2020, October 12). Eight Persistent COVID-19 Myths and Why People Believe Them. Scientific American. https://www.scientificamerican.com/article/eight-persistent-covid-19-myths-and-why-people-believe-them/

8. Shahsavari, S., Holur, P., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news. *arXiv preprint arXiv:2004.13783*.

9. Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3-35.

10. Strange conspiracy theories: from 5G to Meghan Markle. (2020, May 22). The Week. https://www.theweek.co.uk/conspiracy-theories/62926/the-strangest-conspiracy-theories-from-meghan-markle-to-paul-mccartney

11. Strange conspiracy theories: from 5G to Meghan Markle. (2020, May 22). The Week. https://www.theweek.co.uk/conspiracy-theories/62926/the-strangest-conspiracy-theories-from-meghan-markle-to-paul-mccartney

12. Olito, F. (2020, November 11). Insider. https://www.insider.com/popular-conspiracy-theories-united-states-2019-5

13. van Prooijen, J. W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on psychological science*, *13*(6), 770-788.

14. van Prooijen, J. W., & Van Vugt, M. (2018). Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on psychological science*, *13*(6), 770-788.

15. Griffin, A. (2020, November 1910. Independent. https://www.independent.co.uk/news/uk/home-news/princess-diana-death-conspiracy-theories-b1746545.html

16. Identifying Conspiracy Theories. (n.d.). European Commission. Retrieved May, 5, 2021, from https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/fighting-disinformation/identifying-conspiracy-theories_en

17. Wood, M. J. (2018). Propagating and debunking conspiracy theories on Twitter during the 2015–2016 Zika virus outbreak. *Cyberpsychology, behavior, and social networking*, *21*(8), 485-490.

18. Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots?. *First Monday*.

19. Mahir, E. M., Akhter, S., & Huq, M. R. (2019, June). Detecting Fake News using Machine Learning and Deep Learning Algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

20. Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Political Psychology*, *40*, 3-35.

21. Shahsavari, S., Holur, P., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news. *arXiv preprint arXiv:2004.13783*.

22. Serrano, J. C. M., Papakyriakopoulos, O., & Hegelich, S. (2020, July). NLP-based feature extraction for the detection of COVID-19 misinformation videos on Youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

23. Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference* (pp. 372-378). IEEE.

24. Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference* (pp. 372-378). IEEE.

25. Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference* (pp. 372-378). IEEE.

26. Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference* (pp. 372-378). IEEE.

27. Mahir, E. M., Akhter, S., & Huq, M. R. (2019, June). Detecting Fake News using Machine Learning and Deep Learning Algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

28. Mahir, E. M., Akhter, S., & Huq, M. R. (2019, June). Detecting Fake News using Machine Learning and Deep Learning Algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

29. Mahir, E. M., Akhter, S., & Huq, M. R. (2019, June). Detecting Fake News using Machine Learning and Deep Learning Algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

30. Mahir, E. M., Akhter, S., & Huq, M. R. (2019, June). Detecting Fake News using Machine Learning and Deep Learning Algorithms. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

31. Novakovic, J., & Rankov, S. (2011). Classification performance using principal component analysis and different values of the ratio R. *International Journal of Computers Communications & Control*, *6*(2), 317-327.

32. Novakovic, J., & Rankov, S. (2011). Classification performance using principal component analysis and different values of the ratio R. *International Journal of Computers Communications & Control*, *6*(2), 317-327.

.

33. Zhang, Yin & Jin, Rong & Zhou, Zhi-Hua. (2010). Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 1. 43-52. 10.1007/s13042-010-0001-0.

34. Mujtaba, H. (2020, September 11). An Introduction to Bag of Words (BoW) | What is Bag of Words? Great Learning. https://www.mygreatlearning.com/blog/bag-of-words/#sh6

35. Zheng, A. & Casari, A. (2015). Chapter 4. The Effects of Feature Scaling: From Bag-of-Words to Tf-Idf. O'Reilly. https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/ch04.html

36. Koehrsen, W. (2018, March 3. Beyond Accuracy: Precision and Recall. Onwards Data Science. https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

37. Koehrsen, W. (2018, March 3). Beyond Accuracy: Precision and Recall. Towards Data Science. https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

38. Gandhi, R. (2018, July 3). Support Vector Machine — Introduction to Machine Learning Algorithms. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

39. Gandhi, R. (2018, July 3). Support Vector Machine — Introduction to Machine Learning Algorithms. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47