ABSTRACT

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. Fake news has become such a big problem because it possesses the ability to influence and eventually change opinions of the readers and viewers and also impacts the way in which people respond to an authenticated news article. It also has political influence, can encourage mistrust in legitimate media outlet, influence financial markets and damage an individual's reputation. We aim to develop a model using machine learning and Natural Language Processing (NLP) techniques to determine whether a news is fake or real.

This report examines existing 'Fake News' detection systems and highlights the drawbacks of these systems, particularly in relation to the use of classification algorithm that should be chosen and how they all should be combined to get the maximum accuracy. Furthermore, proposed implementations of 'Fake News' detection system using machine learning technology has been researched in order to identify how Term Frequency-Inverse Document Frequency (TF-IDF), Parts of Speech (PoS) and semantic analysis can be combined to attain maximum accuracy for the classification model.

ACKNOWLEDGMENTS

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and our college. We would like to extend our sincere thanks to all of them.

We are highly indebted to Professor Nalina V. for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project.

We would like to express our gratitude towards our parents for their kind cooperation and encouragement which helped us in completion of this project.

Our thanks and appreciations also go to our colleagues in developing the project and people who have willingly helped us out with their abilities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Overview

In this era of Information Technology when we say data is the new oil, although it has made a revolutionary change in our lives but at the same time it has also posed some serious threat to civilisation. This digital age or the New Media Age has not only made us realise our dream of making the world a global village but has also made the long distances seem much shorter and removed barriers with every new discovery. The technology has become an essential part of our lives and now we are so dependent on it that we cannot imagine a day without our modern age tools. We are no longer only a citizen of our country or the city we live in but have become a global citizen as the amount of time it takes for a local news to reach us, is nearly the same or sometimes even more than what it takes to reach from a far-far land somewhere in the world. We are now on a network as big as the world called the 'Internet'. Today information is anyone's principle asset and the sources to get this information are huge in number at the same time volume of information available is also enormous.

The IT revolution has led to enormous changes in each and every field that touches human lives. We can find it in the service sector, manufacturing sector and now in the agriculture sector also. The service providers, be it big players like railways and other national transporters or some small restaurant in the urban lanes, IT has come out to be everyone's friend. It has given a powerful device in every hand with some 'super powers' being a click away from such devices, but as Mr. Winston Churchill once said, "Where there is great power there is great responsibility". The project tries to answer one such responsibility that has come on the shoulders of every concerned and vigilant citizen of the

world as a result of the IT revolution, the technological advancements and reachability of the internet to the masses throughout the world.

This report identifies and addresses the issue and menace of ‘Fake News’ that has spread and found its way much deeper into the society and now it is high time it should be addressed. The system has been developed by choosing the weapon on the principle of ‘Diamond cut Diamond’ to tackle this menace, that is software development with the help of modern age Information Technology tools.

1.2 Motivation

The issue of how to tackle the disruption done by ‘Fake News’ has now become a very crucial issue for both the governments around the world and the big technology companies. Before going into the details of ‘Fake News’ we should look at how this problem got its shape over the years in the world. With the invention of the World Wide Web (WWW) by computer scientist Tim Berners-Lee the world of internet took more recognizable form in the 1990s, before this time the information sources were comparatively less in number and thus these recognised and well identified resources would transmit and give out only factual and real news to the general public and if anyone maliciously tried to spread fake news then the propagation speed used to be too slow for it to affect the lives of people or to shape public opinions on any matter. Also, they were handleable easily due to their lesser reach to the masses and slow propagation speed.

After internet got its recognisable shape in and around the 1990s, the world saw a revolution in terms of reachability to any information and the technological advancements that happened parallelly led to spread of internet enabled personal computers and televisions. In India the BBC broadcast started in 1937 and people could watch it only if

they had a satellite or cable connection and television sets, which were obviously very rare at that time. Till 1990s 'Doordarshan' was the only Indian news channel available in India but soon private players started to emerge and as a result by 2005 India had 200 digital channels, which grew to 800 in 2012 which includes 400 news channels alone. Thus, we can see the emergence of the new sources of news for the people and certainly all these sources are being operated by different people, which makes the system more vulnerable to spread of unverified news. The onset or beginning of social media enabled people to control the propagation of information for the first time, as now the news came out of the less interactive news rooms on their television sets.

After the invention of blogging in late 1990s the popularity of social media has increased like never before and sites like MySpace and Linkedin got fame. Facebook and Twitter were available to people across the world by 2005 and it led to social media age where everyone has a profile and preferences mentioned on that profile. WhatsApp, that was launched in India in mid 2010 has become a great source for instant messaging but also it has now become a major threat for 'Fake News' menace, with more than 10% of total WhatsApp users alone being from India. This picture provides an insight as to how the propaganda stories are really making people to commit any crime or be manipulated according to that.

Not only the people with less access to modern education but technically aware people are also being trapped in it, as even those people have less time to analyze the information they are receiving and sharing, and examine it for its authenticity. We cannot leave the responsibility to tackle the problem of 'Fake News' on people only. Thus, a need of a modern-day tool to help people in deciding authenticity of the information they are sharing is felt, something this project aims at doing.

The scene today in 2019 is such that India has the cheapest mobile data in the world as per March, 2019 with 1GB being just Rs.18.5, with such cheap and affordable data available on 4G services each and every section of the society has access to the internet and thus, to the information being transmitted over that through various mediums but most importantly the social media. Therefore, the solution developed should be such that it should be easily deployed to people in different sections of the society.

1.3 Knowing the Stakeholders

Before designing the solution for the problem we should try to understand how the 'Fake News' spreads, who is their target, what is their objective and the objects because once we know the source and motives of the 'Fake News' we can now think about the solutions that can be provided by seeing the general trends in their origination source, the language used to write the news and the sentiments involved in the message. Upon examining fake news from across the world, it was observed that most of the fake news were created to form public opinions about political parties and to fulfill one's religious motives. Thus, it seemed possible to use semantic analysis to identify some traits of 'Fake News' in any news article because the grammar used in the articles tried to appeal to the emotional side of the reader rather than providing facts to them. Even the parts of speech used were not evenly distributed, generally it was found that a lot of adjectives and adverbs were used in articles that were fake or unauthorised.

The type of algorithms used by the social networking sites like Twitter, Facebook etc. were analyzed to get an answer as to how the 'Fake News' article get shared to millions of people in all parts of the world without being restricted or cross checked at any level. It was found that big social networking sites work on deep learning algorithms that have

learned to give priority to the contents or shared articles or information that already have a greater engagement or that have already been shared larger number of times in comparison to other less shared posts, but we need to note the fact that the post or 'Fake News' has to be shared already a good number of time in comparison with other posts. This led to conclusion that the spread of 'Fake News' not only depends on technological factors but it is a mixture of both technological and human factors.

It is understandable that according to human nature, we are very likely to share the contents presented to us that address our grievances and is inclined with our belief system like religious or political preference, often without even examining the source of that content or fact checking the viral piece of information. This means that such inflammatory contents will generate a fast engagement and therefore is likely to be shared more by the public or viewers. It is only after this human interaction that the technology's role comes into the picture as the algorithms there are pre-designed to show the same inflammatory pre-shared content to other viewers and thus worsening the situation even further. This creates a scenario where any one visiting the social media is exposed to a complete collection of unauthorised and invective news. It was also felt that there is need for a proper definition as to what is 'Fake News' because we know that if a problem is well-defined it might include its solution in the problem itself.

Thus, 'Fake News' can be defined as a piece of information which can be a news article or any content in audio/video or written format, that is created intentionally to misguide, influence or establish a propaganda among its readers, listeners or viewers. It can spread through traditional news media or social media. The problem of 'Fake News' is severe as it tries to be very similar to any authenticated news information from a reputable source and thus, people tend to believe them and fall in the trap. The continuous involvement of governments around the world in the dirty game of 'Fake News' to win

elections or propagate their political motives has further worsen the situation as the law makers are themselves suspected to be involved. Therefore, the technology and public awareness seems the only forward to reduce vulnerability towards this menace. A revolutionary step is needed to make the citizens to our country 'Technologically Literate' so that people on their own discretion with the help of technology can decide between an authorised and unauthorised news content.

1.4 The Technology Stack Involved

This report provides the technological solution towards detection of 'Fake News', as this solution of detecting the fake news will work at the user end and as checking each and every source of content generation in the vast internet network, where everything can be posted remotely and anonymously would be very difficult. Governments around the world will find it difficult in terms of their jurisdiction to regulate the sources of 'Fake News' themselves and it would require a large amount of resources also. The presented system incorporates the field of 'Machine Learning' to design the solution.

Machine learning is basically making our systems or machines learn to do tasks without having to provide them explicit command for execution of that task. Thus, roughly our solution can be thought of making the viewers devices to be able to decide on their own the contents on the systems to be real or fake and hence warn the viewer. Most of the people working with huge data, which is a common phenomenon nowadays, have recognised the value of machine learning for their work and extensively use it also. Today machine learning is being used in various sectors to help the people, like it is being used in the Financial Services where banks are extensively using machine learning to perform data analytics to improve their service experience and to design the policies for their customers.

Banking frauds can also be prevented by implementing machine learning algorithms. Similarly, in the HealthCare sector machine learning has provided some very efficient lifesaving tools like sensors and wearable devices that monitor patient's health condition.

In transportation Sector the use of machine learning has changed the way people travel across the world. The transporters can now decide the most viable and demanding travel route and provide services accordingly to the commuters. The Governments across the world are using machine learning techniques to draft public health policies and perform impact analysis of their citizens and also to identify theft and analyze the aspects of national security. The bottom line is that in our solution we can learn the ways machine learning is being used to perform functions mentioned above and analyse the algorithms being used by them because the same algorithms can be scaled up or down for providing solution for our problem.

Reviewing state of the art work on an existing problem is very necessary for understanding the challenges of designing the solution and also get different point of views on the problem. In these reviews we find algorithms with an improved capability for classification and the techniques we can use to reduce overfitting in the classification model. Here different classification algorithms and their combination will be used. In this problem the output that we want for any news article is variable in the form of categories, i.e. either 'REAL' or 'FAKE' and the classification model would look at the input data and try to predict labels for desired result.

As the nature of the problem requires examining the news articles collected from different sources like websites and social media platforms, it is quite sure that the model will require looking into and analysing most of human written material in any of the human readable/writable language, followed by a classification of them being either 'REAL' or 'FAKE'. This task will be performed with the help of Natural Language Processing (NLP)

which is the process of manipulating and drawing some valuable results out of a natural language, i.e. a language used by humans to communicate and publish their thoughts and ideas across the world. The design of systems that deal with NLP input and output has always been a complex task and a task full of risks of computational errors primarily because however easy and understandable language with a wonderful grammar be present there, but we humans lack to formally understand and design the rules that govern the language. This inability makes it difficult for software engineers like us to make our machines or systems to analyze and get a result out of the natural language that is being processed.

The problem of classifying ‘Fake News’ posed a problem in terms of deciding whether to use rule based NLP or statistical NLP technique, as the model examines texts in which there has to be a proper format for writing the news article otherwise it has a high probability of being fake, this requires few rules regarding the way in which the natural language is put or formulated in the article. Statistical NLP technique should be used as it is needed to know the frequencies of the text segments in an article or in the whole document because fake and real news have different tendencies of frequencies of words as the fake news have to propagate a political, religious or economic agenda and they need to repeat words in order to validate their point.

As the problem talks about spreading propaganda about grievances of people and their belief system and thus influence them to make a particular political or economic decision for the benefit of people spreading fake news, it has to deal with the sentiment of people or we can say that ‘Fake News’ article mostly plays with people’s sentiments in order to manipulate them to take certain decisions. Thus, sensing or examining the sentiments of the news article that are being examined for classification becomes very

important and without doing a sentiment analysis the nature of the news article cannot be fully judged and concluded.

Sentiment analysis uses different tools like NLP, text analysis, computational linguistics etc. to get a sense of or extract the polarity of a given text. It is already being widely used in the service sector for customer satisfaction like in review systems etc. It tells whether text to be examined is positive, negative or neutral in nature by examining its linguistic features. In our problem, it is possible that the fake news, since it is trying to polarise the mind of the readers towards one side and make them to take action based on such bias which can be on religious or political basis, be representing a lot of polarised sentiments than an authorised real news, since the genuine news readers and writers tend to be unbiased and just want to deliver the information and don't engage in shaping public opinion. So, analysing the sentiments of the text becomes very important.

1.5 Scope

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 respond to real news. It has political influence, can encourage mistrust in legitimate media outlet, influence financial markets and damage an individual’s reputation. The system presented in this report has to develop a model using machine learning and NLP techniques to determine whether a news is fake or real.

Chapter 2

LITERATURE SURVEY

Automatic deception detection [1], in this publication the authors discussed two methods for Fake News detection. The first one is linguistic approach; this discusses the various syntactic and semantic features that are useful in deception detection. Network approach depends on querying existing knowledge networks, or publicly available structured data, such as DBpedia ontology, or the Google Relation Extraction Corpus (GREC). They have expressed the need to identify the person or group posting news on any social media platform is important for having trust on the article.

Evaluation of classification algorithms [2], in this research work the authors have explored various Natural Language Processing techniques for the detection of fake news detection. In this paper, the authors have performed comparison and by the means of comparison they have tried to get a model that best suites the classification mechanism. They have used three different factors of comparison:

1. TF-IDF using bi-gram frequency,
2. syntactical structure frequency (probabilistic context free grammars, or PCFGs)

They have used

1. TF-IDF and bi-gram frequency,
2. Syntactical structure frequency.

To get the final model that will be used for detection using examination of text based on NLP.

Data Mining Perspective of Fake news detection [3], in this paper the authors have emphasized that they will be using sentimental features as well as various lexicon features of the news article provided as input to the machine learning algorithm. The sentimental

features can be best understood and learned from the data mining tools, i.e by learning from a collection of very carefully collected and parsed data using modern data mining tools.

Using Syntactic Stylometry [4], in this state-of-art work the authors have used Support Vector Machining (SVM) classifier and used syntactic stylometry for deception detection rather than just shallow lexico-syntactic patterns. They used the Context Free Grammar parse trees to derive features which were used for classification. The datasets were customer reviews on service provider's platform some of them were genuine reviews and others were generated by Amazon Mechanical Turk.

Using Network and Linguistic Approach for classification [5], this state-of-art work highlights two approaches for detecting fake news. The way that the authors have used is the Linguistic Approach, the goal of this method is used to find "leakage" in the language or the text provided as input. This includes examining data representation, deep syntax and semantic analysis etc. The second approach used was Network Approach which involves using network properties which complement the methods described in Linguistic approach for fake news detection, this includes Linked data analysis and social network behaviour analysis.

Deception detection in Speech [6], in this research work the authors have highlighted the need for deception detection methods in children's speech for legal purposes, the methods used in this approach can also be used to deception detection in fake news analysis as they used random forest algorithm as one of the classifying algorithms. They searched for deception in transcribed, typed, and handwritten text by identifying features of linguistic style such as the use of personal pronouns and exclusive words (e.g., but, except, without).

Using TF-IDF for analysing Comments [7], this state-of-art work incorporates a widely used term weighting technique used for text pre-processing called TF-IDF. The text

to be classified was pre-processed using TF-IDF weight after that various interdependencies of comments were categorised as follows comment-to-commenter, comment-to-social network, comment-to-document and comment-to-comment, these dependency structure between observation in study, thus introduced a statistical bias, this bias, if ignored, can manifest in a non-robust method at best, and can lead to an entirely wrong conclusion as worst. They have proposed that by bringing some changes in the TF-IDF factor how they can improve the performances of the developed model. The proposed change in the factor can be incorporated in our model to improve performance.

Facebook Operations Information [8], in this official publication of social media giant Facebook they have acknowledged the risks and responsibilities that they have towards the society. They have aimed at improving the public opinions and stop them from manipulating public opinion. The authors also talk about and present a detailed report on the U.S presidential elections of 2016, which makes us realize the menace fake news has created around us.

Using NLP to classify Fake News [9], in this the authors have used quoted content and the forward and backward attribution span of the quoted text in a document and perform NLP on the same in order to classify the document as real or fake. They found that, of the 60 documents reviewed, it was identified that the preponderance of the false content documents (28 of the 30 reviewed documents) that included quotes either lacked proper attribution or attributed quotes to non-named entities to assert a fact. They build an attribution classifier and the resulting binary classification label was based on the presence of learned source and cue information inside the attribution spans. To identify a source, the custom classifier searched for named-entities or persons or organizations that could be attributed as having made a quote using named-entity recognition methods. Cue

identification is based on learning associated cueing verbs or cue information contained inside the training set.

An improvement in Naive Bayes classifier for review analysis [10], in the presented research paper the authors give a framework for analysing the customer reviews in a restaurant and in order to do so they suggest some changes in Naive Bayes algorithm that would lead to improvement. The paper very clearly produces the comparison between the old and new algorithm results and hence proves their point of improving performance by improving the algorithm itself. They collected restaurant reviews in order to improve the problems in the existing research, and a restaurant senti-lexicon was developed after manually analysing the collected data. The model searches for the sentiment expressing words or phrases in the review text provided as input to the machine and these sentiment or expression rich words decide the positiveness or negative nature of the review and eventually after aggregation of all such values obtained for all the keywords or phrases a final classification is provided.

Stopping click baits on news media [11], in this the authors have attempted to automatically detect clickbait and then build a browser extension which warns the readers of different media sites about the possibility of being baited by such headlines. This approach is necessary and can be very useful for preventing users from accessing false news, but it also poses some challenges as news articles are not as easy to classify as the click bait, therefore classification itself will remain a serious concern in this model. For developing their model, they have used sentence structure features like

1. sentence length,
2. word length
3. syntactic dependencies and stop words,

4. hyperbolic and common phrases such as (internet slang, hyperbolic words, punctuation patterns and common bait phrases)

Using these features for differentiating between clickbait and non-clickbait headlines they attempted to automatically detect clickbait and then build a browser extension which warns the readers of different media sites about the possibility of being baited by such headlines. They used SVM, Decision Tree and Random Forest classification algorithms for the classification.

Rumor detection on social media [12], it emphasizes that how easy it is to spread fake news at a fast pace with the increase in technology and social media. The paper proposes an idea to identify newly emerging rumours by learning from historical data with fake news. The framework, Crosstopic Emerging Rumor detection (CERT), identified and utilized pattern such as curiosity, skepticism, and astonishment from prior labelled data to help recognize new fake data. Their current work was only done using text related to fake data. They mentioned future aspects of the project by taking videos and images as fake data and then making a framework to detect fake news

Online news story life cycle analysis [13], in which they took help of the journey that any news article goes through after being posted online and examined their whole life cycle. This was necessary to do in order to predict future visitation patterns early and accurately. They used data for developing their framework from a very well-established news organisation called Al Jazeera and this data was collected from their social media pages like Facebook and Twitter. They made a framework which focused on different attributes of online posts and integrated different types of interactions of users with the news article which included social media reactions, searches and referrals. They not only focused on a single type of news but examined different news types like breaking news,

in-depth news etc. to make their final framework to understand the reaction of readers to the online posts.

Methods to separate Fact from Fiction [14], in the state-of art works researchers have tried to come up with a framework to classify suspicious or verified news posts and predict which type of suspicious news which can be a satire, hoaxes, click bait and propaganda. Thus, they have nearly covered all major objectives of the fake news propagator. They came up with neural network models which were trained on data from twitter and social media interactions instead of only linguistic features and found that social network interactions outperform lexical models. They performed the classification on the data collected from Twitter for a period of 2 weeks. They used network specific and features to enhance the performance of classifier. They also used text classifications like Long Short-Term Memory and Convolution Neural Networks.

Detecting frauds in Writing style over the internet [15], in this research work the authors have used linguistic features and stylistic features to recognize deceptive content. If we initially identify some suspected written content over the internet then we can attach a warning with that content for users to know that the content can be dangerous or unverified and hence, user discretion is required. They performed analysis on the Brennan-Greenstadt adversarial dataset and a similar dataset collected using Amazon Mechanical Turk (AMT).

They used features like:

1. quantity (number of syllables, number of words),
2. grammatical complexity,
3. uncertainty and specificity,
4. expressiveness to classify data.

They used Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO). Testing was done in the WEKA tool. The paper focused on materials that are made by copying the style of other authors to fool readers because such content can very easily deceive people who read it and hence even the uses of viewers how remain vigilant towards unauthorised news also come in the trap. This paper comes up with a framework to detect those texts.

Chapter 3

AIM

3.1 Aim

As the epidemic of fake news is growing, it is becoming increasingly important to curb this problem. Deliberately misleading news published under the guise of real news is a global problem which affects voting decisions, public conception and judgement. Fake news impacts people in several ways including capacity to shape regional and national dialogue, cause hoaxes and hurts businesses. Significant effect of fake news can be seen in the Indian General Elections 2019 and US Presidential Elections 2016. With the surge in news emerging from several sources, organization, it is difficult to confirm news utilizing conventional checkers. The propaganda spread through fake news can impact individuals' attitude and the way they conduct themselves with other citizens. Fake news can also be spread by rivals to get ahead in the competition by spreading phony news about sales and services of a particular organisation. False news can also disturb peace in society and instill fear in people if fake news like natural calamity or bomb attacks are spread. Systems have to be built to safeguard people from situations that arise from fake news.

3.2 Problem Statement

From the above discussion it can be noted that how the fake news has become a big menace to the people throughout the world. No doubt that the user's or viewers' discretion is required for complete protection from the fake news but using technology it is possible to design and implement a model that will detect the fake content being circulated over the

internet or any news media. This system should be able to detect with an adequate confidence whether the news article is fake or not.

3.3 Objectives

The major objectives of the project are threefold, as follows:

1. Using linguistic cues to develop a machine learning based model for accurately determining authenticity of the given news.
2. Demonstrate successful use of classification algorithms to classify the feature vector generated from dataset into real or fake.
3. To get high accuracy to determine whether a news is fake or true.

Chapter 4

METHODOLOGY

4.1 Data Collection

The most important part of solving a machine learning problem is finding data. The classifier made is as good as the data that is selected in this step. Datasets used in this project were taken from Kaggle.com which lets users discover and use datasets available, build models in a web-based data-science environment and gives a chance to learn from other data scientists and machine learning engineers. Data collected was combined into one document to be used for training and testing the classifier. The labels were changed to 'Fake' and 'Real' for all the datasets. Quality of data after this step is not good enough to be used for the project. Hence, data preprocessing is required.

4.2 Data Preprocessing

Data Preprocessing is used in machine learning to change raw data into an understandable format. Data obtained data collection is not suitable to be used for the classifier. This step includes removing duplicate values as several datasets were combined into one which resulted in some duplicate values to be present in the final dataset. Duplicate values should be removed as they increase the size of dataset and can create bias for classification algorithms like Bayes. Null values are also removed from the dataset as they negatively affect the performance and accuracy of any machine learning model. Data points with character length less than 10 are removed as they do not contribute much to the classifier and can make the classification less accurate. Punctuations and non-English alphabets are removed from the dataset. All the words are converted to lowercase. This is done so that the frequency of each word is counted properly. This step is necessary since

words like ‘the’ and ‘The’ will be looked at differently by the machine. Documents use different forms of the same word, such as *playing*, *plays*, *played* and so on. All the words are converted to its base form to get a thorough count of each word to get a gist of what the article is talking about. Dataset obtained after applying all the steps is used for training and testing the classifier.

4.3 Generating Feature Vectors

4.3.1 Bag of Words

Bag of words is a method for representing text in a way that can be processed by machine learning algorithms. It extracts features from textual content. There are 2 major factors which are considered in this type of representation: vocabulary of the words and count of presence of the words in each document. The order of the words or the structure of the sentence is not taken into account, only the fact whether the word is present in the document or not is examined. In this project we have used a refined approached to create vocabulary of combination of two words. This sophisticated approach allows the model to capture more context and meaning from the document. Each word token is called a “gram”. Creating two-word tokens is called a bigram model. We could also consider creating trigram tokens which are a three-word series of words to extract more interpretation of the text, but this could lead to a phenomenon called the curse of dimensionality.

4.3.2 TF-IDF

TF-IDF stands for term frequency-inverse document frequency. TF-IDF is a metric that shows how important a word of a document is when compared to all the documents. For example, if ‘NASA’ is mentioned numerous times in one document and not so often in

others, it probably means that it is of relevance to that document. TF-IDF is calculated by multiplying 2 different parts: The term frequency of a word in a document is calculated by getting the frequency of a word in a particular document and the inverse document frequency is calculated by getting a measure of how prominent the word is in all the documents. Most repeated words are given less weight and less frequent terms are given more weight. If the word is very common and appears in many documents, the measure will approach 0, otherwise it will be close to 1. TF-IDF is used as machine learning algorithms deal with numbers that a machine can understand rather than text. TF-IDF is used to convert text into number so that machine can work on it to make an algorithm. TF-IDF is put as input into the classification algorithms. Each word is considered a token in TF-IDF. For generating the feature vector, TF-IDF values of the bigrams are calculated and represented in TF-IDF vectors.

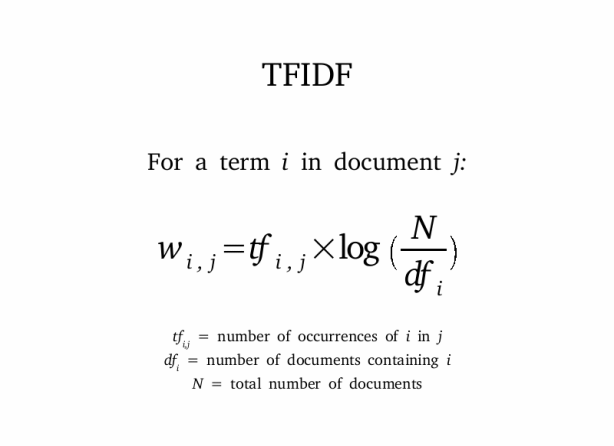

Fig. 4.1 Formula for calculating TF-IDF

4.3.3 Syntactical Analysis

In any language, syntax and structure for the most part go in conjunction, where a set of rules guide the manner in which words are combined into phases. Part of Speech (POS) tagging is a typical parsing technique used for understanding text syntax which classifies

words into their parts of speech. POS are definite lexical categories such as nouns, verbs, adjectives, adverbs and so on, to which words are allocated, in light of their syntactic context and role. POS tags were generated using the Spacy library. The POS features were encoded using TF-IDF vectorizer and further strengthened using bigrams.

4.3.4 Semantic Analysis

The semantic analysis is done by reading the whole document in content to get an idea of what the text talks about. It recognizes each word or token and classifies them into semantic categories applicable to psychological processes. It breaks down the context in the surrounding text and it also examines the text structure to precisely summarize the proper meaning of words that have more than one definition. Semantic technology works with the logical structure of documents to get the most important element in document and comprehend the theme of the document. Before semantic analysis, bag of words was used to get the most recurring words in large size of data and then try to figure out what the article talks about by trying to understand why that word occurs the most and what it means. Comprehending this can be a little difficult when words are taken out of context and a word can mean multiple things. Empath is a popular open source tool that can generate semantic categories. It has over 200 built-in categories, such as *government*, *technology* or even emotional tones such as *anger*, *sadness* and so forth.

4.3.5 Combining features

We considered three methods for generating the feature vectors:

1. TF-IDF bigram vector of the text.
2. Feature vector generated by syntax analysis of the text.

3. Feature vector generated by semantic analysis of the text.

After generating these features and their individual feature vector, the combination of all these features to form the final vector is done, on which classification is performed. The combination of the individual feature vectors is performed by assigning weights to each vector and then taking a weighted combination of all the features vector to generate the final feature vector. If x is the weight corresponding to the first feature vector, y for the second and then the weight of the third feature vector will be $1-x-y$. The final feature vector will be a linear combination of these feature multiplied by their corresponding weights.

4.4 Classification

Three different kinds of classifiers were used to compare the accuracy of each classifier when different factors are used.

4.4.1 Naive Bayes Classifier

A Naive Bayes classifier is a probabilistic machine learning model. The model assumes that each feature has an independent and equal contribution to the classifier. Bayes' Theorem finds the probability of a feature given the probability of another feature that is already present. Each feature is assumed to have equal weight.

4.4.2 Random Forest Algorithm

Decision trees are the building blocks of the random forest model. Random forest consists of a large number of individual decision trees that operate as an ensemble. Every decision

tree predicts a class label and the class label with the most votes becomes the suitable model's prediction.

4.4.3 Gradient Boosting Algorithm

Gradient boosting is an example of boosting ensemble model. Boosting is a technique in which the predictors are made sequentially instead of independently. Boosting tries to convert a weak learner to become better. It learns from its errors and tries to reduce the error. Prediction model obtained by the gradient boosting algorithm is a group of weak class label prediction models which are usually decision trees.

4.5 Design

4.5.1 High Level Design

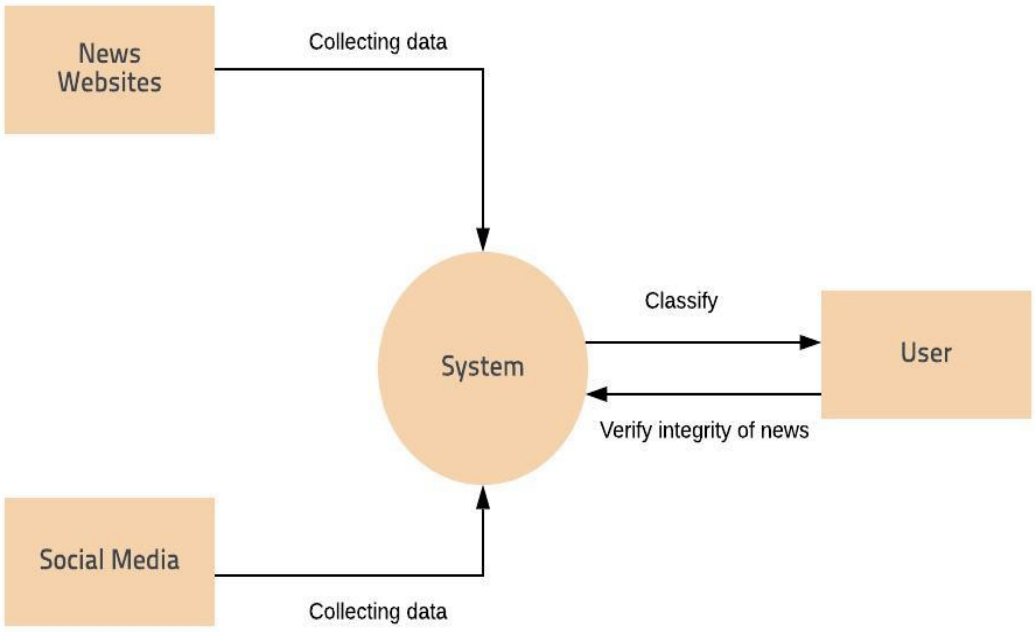

Fig. 4.2 High Level System Design

Figure 4.2 shows the external actors interacting with the system. The external actors are the sources of data and users that use the system. Data to be tested is collected from social media and news websites and the user uses the system to verify the integrity of a news (classify the news article as fake or real).

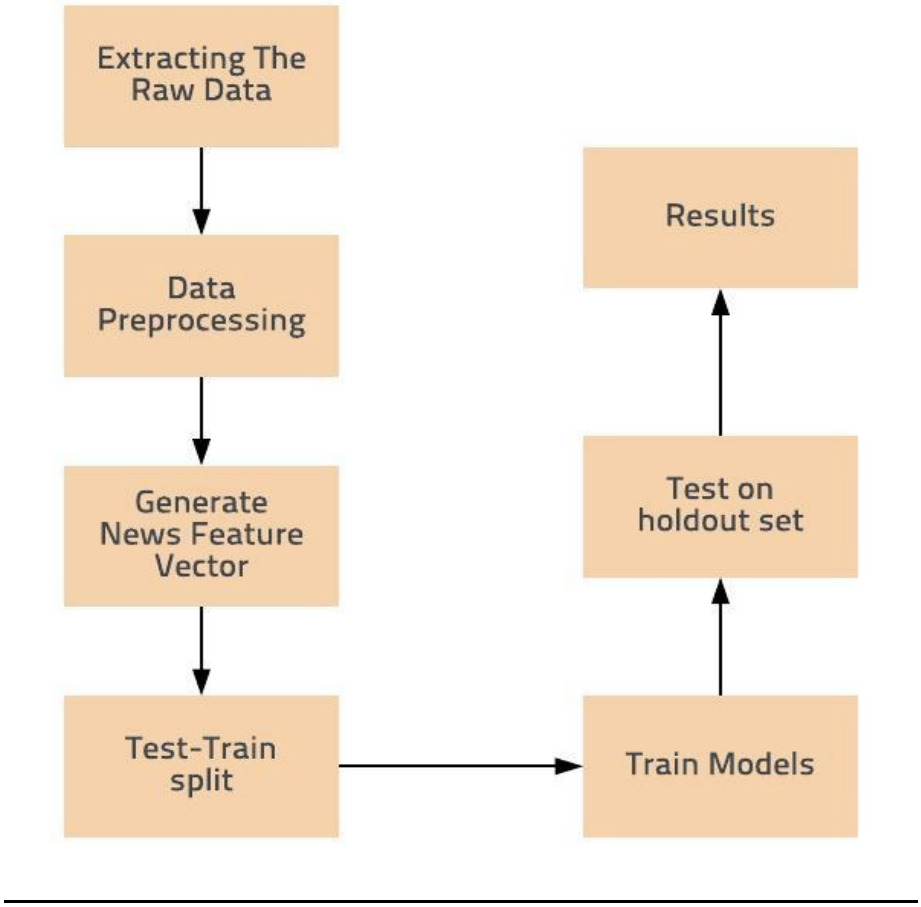

Fig. 4.3 Mid Level System Design

Figure 4.3 shows the different units of our system. Data collected from external sources are saved and processed to get a proper dataset to work with. Depending on the factors chosen, news feature vector is set. Appropriate training models are made according to the news feature vector to acquire high rates of accuracy and precision. The data is divided

into training dataset to train the model and make it learn and test dataset to test the accuracy of the model.

4.5.1 Low Level Design

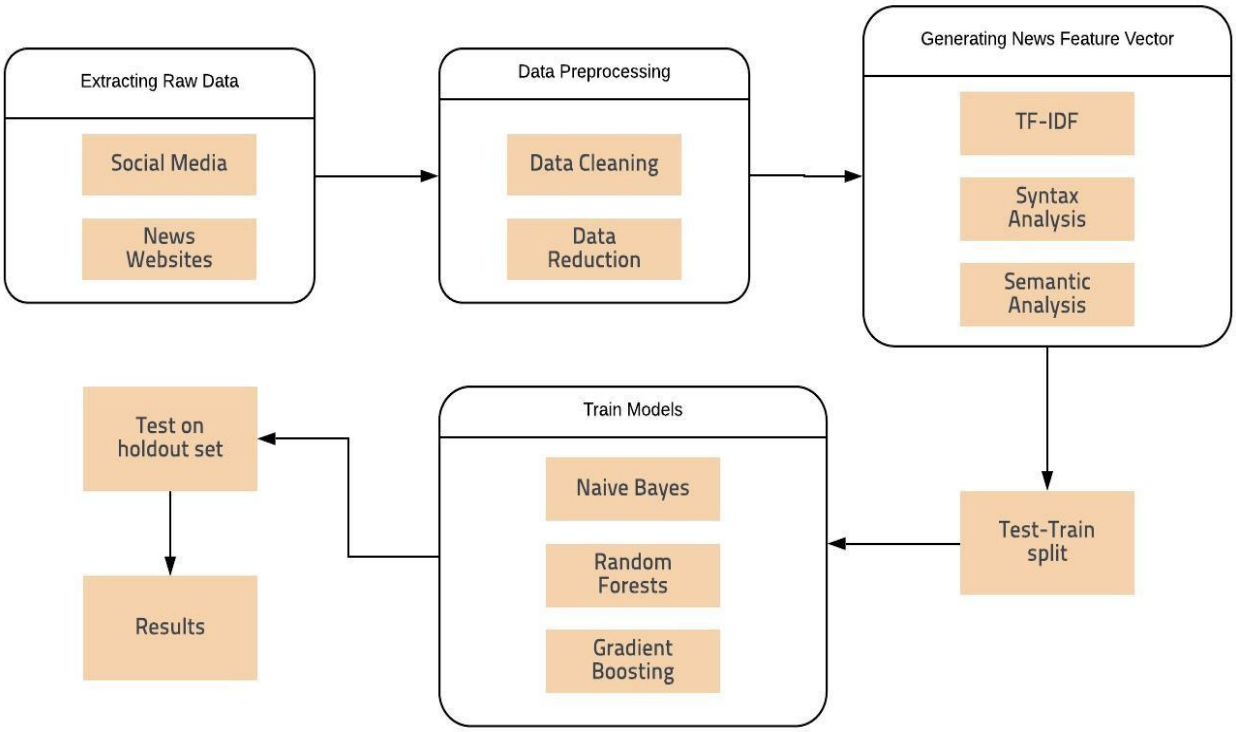

Fig. 4.4 Low Level System Design

Figure 4.4 shows the detailed description of each unit of the system. Data extracted from social media and news websites is preprocessed by removing null and redundant values. Models can be made according to different factors like TF-IDF (Term Frequency-Inverse Document Frequency), which is a statistical measure used to evaluate how important a word is to a document in a collection of documents or syntax and semantic analysis which focuses on the writing style and the emotion conveyed by the document. The data is then

divided into training dataset and test data set so that the appropriate model chosen can be trained to get higher accuracy and then tested to get the results. Different models are taken to compare the accuracy obtained by different models.

4.6 Pseudo Code

4.6.1 Data Preprocessing

```
# Clean text
def clean_text(text):
    # remove punctuation
    text = re.sub('['+string.punctuation+']', '', text)
    text = re.sub(r"[-()\"#/@';:<>{}`+=~|.!?,]", "", text)

    # convert words to lower case and split
    text = text.lower().split()

    # remove stop words
    stops = set(stopwords.words("english"))
    text = [w for w in text if w not in stops]
    text = " ".join(text)

    # remove all non english and numbers etc.
    text = re.sub(r'[^a-zA-Z\s]', u'', text, flags=re.UNICODE)

    # lemmatizing
    text = text.split()
    l = WordNetLemmatizer()
    lemmatized_words = [l.lemmatize(word) for word in text]
    text = " ".join(lemmatized_words)

    return text

# apply clean_text on Text columnxx
df['Clean_Text'] = df['Text'].apply(lambda x: clean_text(x))
df.dropna(inplace=True)
```

4.6.2 TFIDF Vectorization

```
tfidf_vectorizer = TfidfVectorizer(stop_words='english', ngram_range = (2, 2))

tfidf_train = tfidf_vectorizer.fit_transform(X_train)

tfidf_test = tfidf_vectorizer.transform(X_test)

print(tfidf_vectorizer.get_feature_names()[:10])
```

4.6.3 POS Tagging

```
# Generate POS tags
nlp = spacy.load('en')

pos_tags_column = []

for text in df['Text']:
    pos_tags = []
    doc = nlp(text)
    for token in doc:
        pos_tags.append(token.pos_)
    all_pos_tags = ' '.join(pos_tags)
    pos_tags_column.append(all_pos_tags)

df['Text_POS'] = pos_tags_column

df.head()
```

4.6.4 Generating Semantic Category Scores

```
# Getting semantic categories scores
lexicon = Empath()
semantic = []
count = 0

for article in df['Text']:
    d = lexicon.analyze(article, normalize=False)
    x = []
    for key, value in d.items():
        x.append(value)
    x = np.asarray(x)
    semantic.append(x)
df['Semantic'] = semantic
print(df['Semantic'].head())
```

4.6.5 Generating Semantic Classes

```
# Generating semantic classes from semantic score frequency
sem = []
for i in range(df.shape[0]):
    a = []
    for j in range(len(semantic[0])):
        for k in range(int(semantic[i][j])):
            a.append(categories[j])
    b = " ".join(a)
    sem.append(b)
df['Semantics'] = sem

print(df['Semantics'].head())
```

4.6.6 Combining with Weights

```
# setting weights for each feature vector
text_w = 0.35 * 3
pos_w = 0.5 * 3
sem_w = 0.15 * 3

tfidf_train *= text_w
tfidf_test *= text_w
pos_tfidf_train *= pos_w
pos_tfidf_test *= pos_w
sem_tfidf_train *= sem_w
sem_tfidf_train *= sem_w
```

```
# Combining the 3 sparse matrices to form X_train and X_test
# vstack - vertical
# hstack - horizontal
# they are used to combine arrays and make them into one array
diff_n_rows = pos_tfidf_train.shape[0] - tfidf_train.shape[0]
b = sp.vstack((tfidf_train, sp.csr_matrix((diff_n_rows, tfidf_train.shape[1]))))
c = sp.hstack((pos_tfidf_train, b))

diff_n_rows = c.shape[0] - sem_tfidf_train.shape[0]
b = sp.vstack((sem_tfidf_train, sp.csr_matrix((diff_n_rows, sem_tfidf_train.shape[1]))))

# X - train
X_train = sp.hstack((c, b))

diff_n_rows = pos_tfidf_test.shape[0] - tfidf_test.shape[0]
d = sp.vstack((tfidf_test, sp.csr_matrix((diff_n_rows, tfidf_test.shape[1]))))
e = sp.hstack((pos_tfidf_test, d))

diff_n_rows = e.shape[0] - sem_tfidf_test.shape[0]
d = sp.vstack((sem_tfidf_test, sp.csr_matrix((diff_n_rows, sem_tfidf_test.shape[1]))))

# Y - test
X_test = sp.hstack((e, d))
```

Chapter 5

RESULTS AND DISCUSSION

The confusion matrices for the three individual feature vectors used alongside the classification models which gave the highest accuracy are listed from Figure 5.1 to Figure 5.3. The confusion matrices for the combined features vectors used with all the classification models are listen from Figure 5.4 to Figure 5.6.

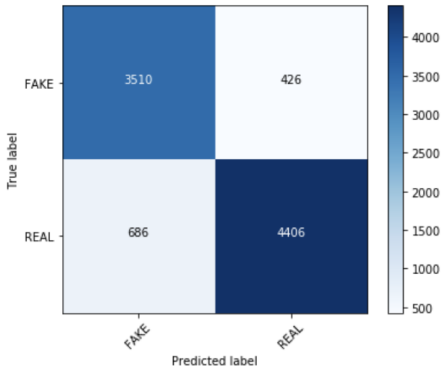

Fig. 5.1 TFIDF RF - 87.6%

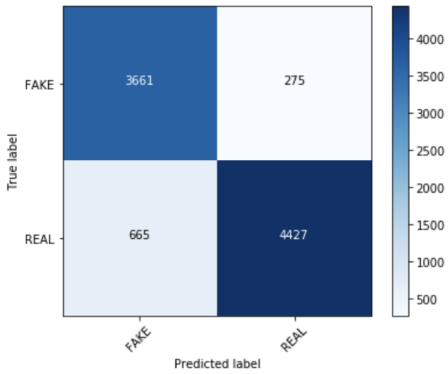

Fig. 5.2 Syntax Analysis GB - 89.6%

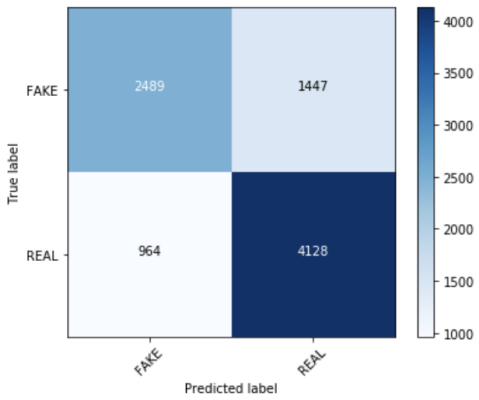

Fig. 5.3 Semantic Analysis GB - 73.3%

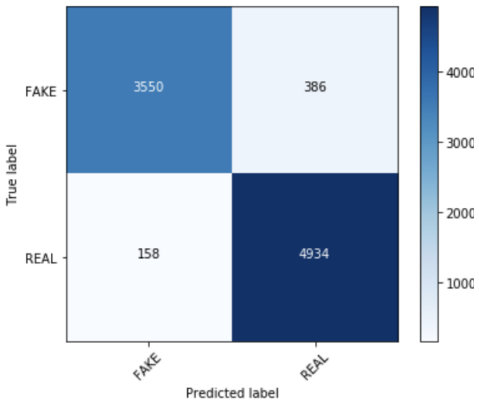

Fig. 5.4 Combined NB - 94.0%

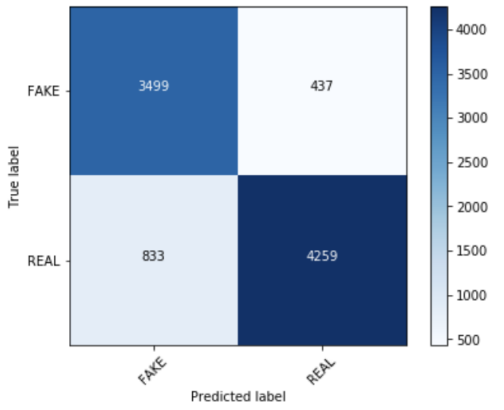

Fig. 5.5 Combined RF - 85.9%

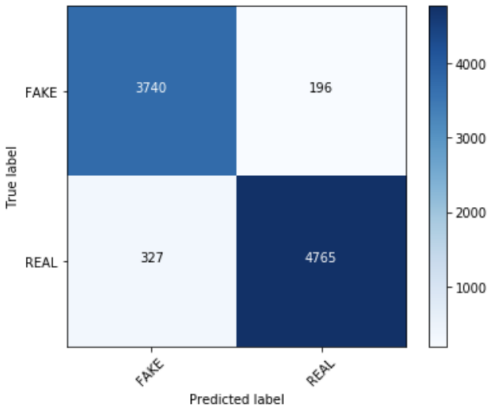

Fig. 5.6 Combined GB - 94.2%

After generating all the feature vectors, they were linearly combined with different weights. The results are compiled in the following table.

Bigrams Vector	Syntax Vector	Semantic Vector	Naïve Bayes	Random Forests	Gradient Boosting
1	0	0	83.9%	87.6%	84.8%
0	1	0	68.3%	87.1%	89.6%
0	0	1	64.1%	69.8%	73.3%
0.35	**0.5**	**0.15**	**94%**	**85.9%**	**94.2%**

Table 5.1 Weights of feature vector and corresponding results

The best accuracy achieved is **94.2%** through Gradient Boosting corresponding to weights 0.35, 0.5 and 0.15 for the feature vectors.

Chapter 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

The ‘Fake News Detection’ model proposed in this report gives a verifiable, reliable, scalable, transparent and dependable approach towards detecting and classifying a news article as ‘REAL’ or ‘FAKE’. The threefold objectives that were kept for development of this classification model were achieved. The classification models were demonstrated with the use of different linguistic cues or linguistic data and features. These linguistic properties helped the classifiers to gain an idea of sentiment of the text or news articles, adequate preprocessing and text cleaning measures were taken to ensure that only the words or phrases relevant in the classification are used as an input to the model for analysis and thus the model does not get misguided by the noise in input data.

The developed model demonstrated the different classification algorithms namely, Naive Bayes Classifier, Random Forest Classifier and the Gradient Boost classification algorithm. How with same algorithms used in combination with different linguistic cues acquired different accuracy rates, tells as why using different algorithms for developing the model was necessary. The use of three different classifiers with each one having a different classifying mechanism, allowed to develop a scalable, verifiable and transparent model.

In order to get the highest possible accuracy in the presented classification model the three different linguistic cues used with three different classification algorithms were combined and that were used only after giving them appropriate weightage points. This enabled development of a model that is not biased or misguided on the basis of any one of the traits that are being examined.

It is hoped that the effective implementation of the proposed machine learning model would eliminate many problems discovered during systems investigation.

6.2 Future Scope

The facts to understand are that, the spread of fake news is not only limited to the textual medium or through the news articles written in a particular language. There have been many instances where video or audio that is being circulated on social media have led to a menace in the society. These audio/video enabled fake news content are very dangerous as it can be spread to people who lack the ability to read or write. Therefore, it seems to have a larger access to the masses. Hence, we need to extend the classification model for the classification of audio/video fake or unauthorised content.

Developing such a model would help the society to fight the menace of fake news in video format and hence the reachability of the classification model will increase. An extension of the present model can be developed to put a check on audio/video enabled unauthorised fake or propaganda content.

REFERENCES

[1] Pérez-Rosas, Verónica, et al. "Automatic detection of fake news." arXiv preprint arXiv:1708.07104 (2017).

[2] Gilda, Shlok. "Evaluating machine learning algorithms for fake news detection." 2017 IEEE 15th Student Conference on Research and Development (SCOReD). IEEE, 2017.

[3] Shu, Kai, et al. "Fake news detection on social media: A data mining perspective." ACM SIGKDD Explorations Newsletter 19.1 (2017): 22-36.

[4] Feng, Song, Ritwik Banerjee, and Yejin Choi. "Syntactic stylometry for deception detection." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, 2012.

[5] Conroy, Niall J., Victoria L. Rubin, and Yimin Chen. "Automatic deception detection: Methods for finding fake news." Proceedings of the Association for Information Science and Technology 52.1 (2015): 1-4.

[6] Yancheva, Maria, and Frank Rudzicz. "Automatic detection of deception in child-produced speech using syntactic complexity features." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013.

[7] Yahav, Inbal, Onn Shehory, and David Schwartz. "Comments Mining With TF-IDF: The Inherent Bias and Its Removal." IEEE Transactions on Knowledge and Data Engineering 31.3 (2018): 437-450.

[8] Weedon, Jen, William Nuland, and Alex Stamos. "Information operations and Facebook." Retrieved from Facebook: https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pdf (2017).

[9] Traylor, Terry, Jeremy Straub, and Nicholas Snell. "Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator." 2019 IEEE 13th International Conference on Semantic Computing (ICSC). IEEE, 2019.

[10] Kang, Hanhoon, Seong Joon Yoo, and Dongil Han. "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews." Expert Systems with Applications 39.5 (2012): 6000-6010.

[11] Chakraborty, Abhijnan, et al. "Stop clickbait: Detecting and preventing clickbaits in online news media." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016.

[12] Wu, Liang, et al. "Gleaning wisdom from the past: Early detection of emerging rumors in social media." Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017.

[13] Castillo, Carlos, et al. "Characterizing the life cycle of online news stories using social media reactions." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014.

[14] Volkova, Svitlana, et al. "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017.

[15] Afroz, Sadia, Michael Brennan, and Rachel Greenstadt. "Detecting hoaxes, frauds, and deception in writing style online." 2012 IEEE Symposium on Security and Privacy. IEEE, 2012.