A Project Report

on

FAKE NEWS DETECTION USING MACHINE LEARNING

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Ms. J. Angelin Blessy
Assistant Professor

Submitted By

Sahil Sourav Sharma(19SCSE1140009)
Tanisha Jaiswal(19SCSE1140036)

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT
OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the project, entitled **"FAKE NEWS DETECTION USING MACHINE LEARNING"** in partial fulfillment of the requirements for the award of the Bachelor of Technology submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during Sept 2021 – Dec 2021, under the supervision of Ms. J. Angelin Blessy , Assistant Professor, Department of Computer Science and Engineering of School of Computing Science and Engineering , Galgotias University, Greater Noida

The matter presented in the project has not been submitted by us for the award of any other degree of this or any other places.

Sahil Sourav Sharma
Tanisha Jaiswal

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Ms. J. Angelin Blessy

Assistant Professor

CERTIFICATE

The Final Project Viva-Voce examination of Sahil Sourav Sharma(19SCSE1140009) and Tanisha Jaiswal(19SCSE1140036) has been held on _________________ and his/her work is recommended for the award of Bachelor of Technology.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December,2021
Place: Greater Noida

Acknowledgement

An endeavor over a long period can be successful with the advice and support of many well-wishers. We take this opportunity to express our gratitude and appreciation to all of them.

We wish to express our sincere thanks and gratitude to our project guide Ms. J. Angelin Blessy, Assistant Professor, Department of Computer Science and Engineering, Galgotias University, for the simulating discussions, in analyzing problems associated with our project work and for guiding us throughout the project. Project meetings were highly informative. We express our warm and sincere thanks for the encouragement, untiring guidance and the confidence she had shown in us. We are immensely indebted for her valuable guidance throughout our project.

We also thank all the staff members of CSE department for their valuable advices. We also thank supporting staff for providing resources as and when required. We extend our gratitude to researches and scholars whose papers and thesis have been utilized in our seminar report.

PROJECT STUDENTS:

SAHIL SOURAV SHARMA	19SCSE1140009
TANISHA JAISWAL	19SCSE1140036

Abstract

In our modern world, the Internet is one of the important inventions and a large number of persons are its users and everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short period. Any user can make a post or spread the news through these online platforms. These platforms do not verify the users or their posts. So, some of the users try to spread fake news through these platforms. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits. This fake news can be propaganda against an individual, society, organization or political party. A human being is unable to detect all this fake news. So, there is a need for machine learning classifiers that can detect this fake news automatically. Fake news detection is an emerging research area that is gaining big interest. It faces however some challenges due to the limited resources such as datasets and processing and analyzing techniques. In this work, we propose a system for Fake news detection that uses machine learning techniques and also check the authenticity of the website publishing the news.

CONTENT

1 INTRODUCTION

2.LITERATURE SURVEY

3. METHODOLOGY

4.DATASET

5.CONCEPTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS, ABBREVIATIONS ANDNOMENCLATURE

LIST OF ABBREVATIONS

SHORT FORM	FULL FORM
CV	Count Vectorizer
W2V	Word 2 Vector
SVM	Support Vector Machine
ANN	Artificial Neural Network

LIST OF SYMBOLS

SYMBOL	MEANING
Σ	Summation(Uppercase Sigma)
A	Alpha
Tanh	Hyperbolic tangent function
Σ	Sigmoid Function(Lowercase Sigma)

CHAPTER 1
INTRODUCTION

1.1 MACHINE LEARNING AND NLP:

1.1.1 MACHINE LEARNING

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relyingon patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "trainingdata", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. "A computer program is said to learn from experience E with respect to some class of tasks Tand performance measure P if its performance at tasks in T, as measured by P, improves with experience E." This is Alan Turing's definition of machine learning.

Deep learning is a class of machine learning algorithms that utilizes a hierarchical level ofartificial neural networks to carry out the process of machine learning. The artificial neuralnetworks are built like the human brain, with neuron nodes connected together like a web.While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach.

The word "deep" in "deep learning" refers to the number of layers through which the datais transformed. More precisely, deep learning systems have a substantial credit assignmentpath (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connections between input and output.

For a feedforward neural network, the depth of the CAPs is that of the network and is thenumber of hidden layers plus one (as the output layer is also parameterized). For recurrent neural networks, in which a signal may propagate through a layer more than once, the CAPdepth is potentially unlimited.

Deep learning architectures such as deep neural networks, deep belief networks, recurrentneural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, audio recognition,social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced resultscomparable to and in some cases superior to human experts.

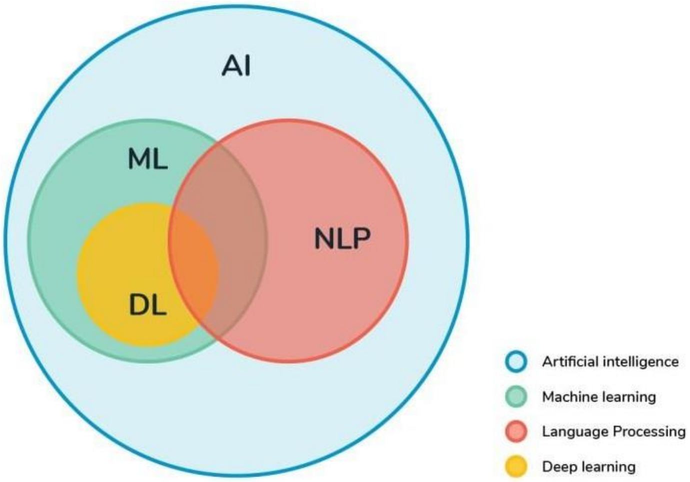


Fig. 1 : Graphical representation of relationship betweenvarious fields in artificial intelligence (source: devopedia.org)

1.1.2 NATURAL LANGUAGE PROCESSING

NLP is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data.

1.1.2.1 STAGES IN NLP

1.1.2.1.1 LEXICAL ANALYSIS

Lexical Analysis involves identifying the structure of words. Lexicon of a language meansthe collection of words and phrases in a language. Lexical analysis is dividing the whole chunkof text into paragraphs, sentences, and words.

1.1.2.1.2 SYNTACTIC ANALYSIS (PARSING)

Syntactic Analysis involves analysis of words in the sentence for grammar and arranging wordsin a manner that shows the relationship among the words. The sentence such as “The school goes to boy” is rejected by English syntactic analyser.

1.1.2.1.3 SEMANTIC ANALYSIS

Semantic Analysis draws the exact meaning or the dictionary meaning from the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the taskdomain. The semantic analyser disregards sentence such as “hot ice-cream”

1.1.2.1.4 DISCOURSE INTEGRATION

The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence. So, in Discourse

Integration gives the meaning based on all the sentences given before it. Eg. Consider the sentence “Water is flowing on the bank of the river” But bank has two meanings: One Financial Institute and Two River of the bank, here System has to consider the second meaning.

1.1.2.1.5 PRAGMATIC ANALYSIS

During this, what was said is re-interpreted on what it actually meant. It involves deriving thoseaspects of language which require real world knowledge.

1.2 MOTIVATION OF WORK

The rise of fake news during the 2016 U.S. Presidential Election highlighted not only the dangers of the effects of fake news but also the challenges presented when attempting to separatefake news from real news. Fake news may be a relatively new term but it is not necessarily a new phenomenon. Fake news has technically been around at least since the appearance and popularity of one-sided, partisan newspapers in the 19th century. However, advances intechnology and the spread of news through different types of media have increased the spread of fake news today. As such, the effects of fake news have increased exponentially in the recentpast and something must be done to prevent this from continuing in the future.

We have identified the three most prevalent motivations for writing fake news and chosen only one as the target for this project as a means to narrow the search in a meaningful way. The firstmotivation for writing fake news, which dates back to the 19th century one-sided partynewspapers, is to influence public opinion. The second, which requires more recent advances in technology, is the use of fake headlines as clickbait to raise money. As such, this paper will focus primarily on fake news as defined by politifact.com, “fabricated content that intentionallymasquerades as news coverage of actual events.” This definition excludes satire, which is intended to be humorous 8 and not deceptive to readers. Most satirical articles come from sources. Satire can already be classified, by machine learning techniques Therefore, our goal isto move beyond these achievements and use machine learning to classify, at least as well as humans, more difficult discrepancies between real and fake news.

The dangerous effects of fake news, as previously defined, are made clear by events in which a man attacked a pizzeria due to a widespread fake news article. This story along with analysisprovide evidence that humans are not very good at detecting fake news, possibly not better thanchance. As such, the question remains whether or not machines can do a better job.

There are two methods by which machines could attempt to solve the fake news problem betterthan humans. The first is that machines are better at detecting and keeping track of statistics than humans, for example it is easier for a machine to detect that the majority of verbs used are“suggests” and “implies” versus, “states” and “proves.” Additionally, machines may be more efficient in surveying a knowledge base to find all relevant articles and answering based on those many different sources. Either of these methods could prove useful in detecting fake news,but we decided to focus on how a machine can solve the fake news problem using supervised learning that extracts feature of the language and content only within the source in question, without utilizing any fact checker or knowledge base. For many fake news detection techniques,a “fake” article published by a trustworthy author through a trustworthy source would not be caught. This approach would combat those “false negative” classifications of fake news. In essence, the task would be equivalent to what a human face when reading a hard copy of a newspaper article, without internet access or outside knowledge of the subject (versus reading something online where he can simply look up relevant sources). The machine, like the humanin the coffee shop, will have only access to the words in the article and must use strategies thatdo not rely on blacklists of authors and sources. The current project involves utilizing machine learning and natural language processing techniques to create a model that can expose documents that are, with 9 high probability, fake news articles. Many of the current automatedapproaches to this problem are centered around a “blacklist” of authors and sources that are known producers of fake news. But, what about when the author is unknown or when fake newsis published through a generally reliable source? In these cases, it is necessary to rely simply onthe content of the news article to make a decision on whether or not it is fake. By collecting examples of both real and fake news and training a model, it should be possible to classify fakenews articles with a certain degree of accuracy. The goal of this project is to find theeffectiveness and limitations of language-based techniques for detection of fake news through the use of machine learning algorithms including but not limited to convolutional neural networks and recurrent neural networks. The outcome of this project should determine how much can be achieved in this task by analyzing patterns contained in the text and blind to outsideinformation about the world.

1.3 PROBLEM STATEMENT

News consumption is a double-edged sword. On the one hand, its low cost, easy access,and rapid dissemination of information lead people to seek out and consume news. It enables thewide spread of "fake news", i.e., low quality news with intentionally false information. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Therefore, fake news detection has recently become an emerging research that is attracting tremendous attention. First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content.

To develop a FAKE NEWS DETECTION system using natural language processing and its accuracy will be tested using machine learning algorithms. The algorithm must be ableto detect fake news in a given scenario.

CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

In the world of rapidly increasing technology, information sharing has become an easytask. There is no doubt that internet has made our lives easier and access to lots of information.This is an evolution in human history, but at the same time it unfocuses the line between true media and maliciously forged media. Today anyone can publish content – credible or not – thatcan be consumed by the world wide web. Sadly, fake news accumulates a great deal of attentionover the internet, especially on social media. People get deceived and don't think twice before circulating such mis-informative pieces to the world. This kind of news vanishes but not withoutdoing the harm it intended to cause. The social media sites like Facebook, Twitter, Whatsapp play a major role in supplying these false news. Many scientists believe that counterfeited newsissue may be addressed by means of machine learning and artificial intelligence.

Various models are used to provide an accuracy range of 60-75%. Which comprises of Naive Bayes classifier,Linguistic features based, Bounded decision tree model, SVM etc. The parameters that are taken in consideration do not yield high accuracy. The motive of this projectis to increase the accuracy of detecting fake news more than the present results that are available.By fabricating this new model which will judge the counterfeit news articles on the basis of certain criteria like spelling mistake, jumbled sentences, punctuation errors , words used .

2.2 REVIEW OF LITERATURE

There are two categories of important researches in automatic classification of real and fakenews up to now:

In the first category, approaches are at conceptual level, distinction among fake news isdone for three types: serious lies (which means news is about wrong and unreal events or information like famous rumors), tricks (e.g. providing wrong information) and comics (e.g. funny news which is an imitation of real news but contain bizarre contents).

In the second category, linguistic approaches and reality considerations techniques are used at a practical level to compare the real and fake contents. Linguistic approaches try to detecttext features like writing styles and contents that can help in distinguishing fake news. The mainidea behind this technique is that linguistic behaviors like using marks, choosing various types of words or adding labels for parts of a lecture are rather unintentional, so they are beyond the author's attention. Therefore, an appropriate intuition and evaluation of using linguistic techniques can reveal hoping results in detecting fake news.

Rubin studied the distinction between the contents of real and comic news via multilingual features, based on a part of comparative news (The Onion, and The Beaverton) andreal news (The Toronto Star and The New York Times) in four areas of civil, science, trade andordinary news. She obtained the best performance of detecting fake news with a set of featuresincluding unrelated, marking and grammar.

Balmas believe that the cooperation of information technology specialists in reducing fake news is very important. In order to deal with fake news, using data mining as one of the techniques has attracted many researchers. In data mining based approaches, data integration

is used in detecting fake news . In the current business world, data are an ever-increasing valuable asset and it is necessary to protect sensitive information from unauthorized people. However, the prevalence of content publishers who are willing to use fake news leads to ignoring such endeavors. Organizations have invested a lot of resources to find effective solutions for dealing with clickbait effects.

2.3 PREVIOUS CONTRIBUTIONS

Shloka gilda presented concept approximately how NLP is relevant to stumble on fake information. They have used time period frequency-inverse record frequency (TFIDF) of bi-grams and probabilistic context free grammar (PCFG) detection. They have examined their dataset over more than one class algorithms to find out the great model. They locate that TF- IDF of bi-grams fed right into a stochastic gradient descent model identifies non-credible resources with an accuracy of 77%.

Mykhailo granik proposed simple technique for fake news detection the usage of naïve Bayes classifier. They used buzzfeed news for getting to know and trying out thenaïve Bayes classifier. The dataset is taken from facebook news publish and completed accuracy upto 74% on test set.

Cody buntain advanced a method for automating fake news detection on twitter. They applied this method to twitter content sourced from buzzfeed's fake news Dataset. Furthermore, leveraging non-professional, crowdsourced people instead of Journalists presents a beneficial and much less costly way to classify proper andfake Memories on twitter rapidly.

Marco L. Della offered a paper which allows us to recognize how social networks andgadget studying (ML) strategies may be used for faux news detection .They have used novel ML fake news detection method and carried out this approach inside a Facebook Messenger chatbot and established it with a actual-world application, acquiring a fake information detection accuracy of 81%.

Shivam B. Parikh aims to present an insight of characterization of news story in the modern diaspora combined with the differential content types of news story and its impact on readers. Subsequently, we dive into existing fake news detection approaches that are heavily based on text- based analysis, and also describe popular fake news datasets. We conclude the paper by identifying 4 key open research challenges that can guide future research. It is a theoretical Approach which gives Illustrations of fake news detection by analysing the psychological factors.

Himank Gupta et. al. [10] gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected

400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of- Words model. 4. They wereable to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%.

2.3 RELATED WORK

2.3.1 SPAM DETECTION

The problem of detecting not-genuine sources of information through content based analysis isconsidered solvable at least in the domain of spam detection [7], spam detection utilizes statistical machine learning techniques to classify text (i.e. tweets [8] or emails) as spam or legitimate. These techniques involve pre-processing of the text, feature extraction (i.e. bag of words), and feature selection based on which features lead to the best performance on a test dataset. Once these features are obtained, they can be classified using Nave Bayes, Support Vector Machines, TF-IDF, or K-nearest neighbors classifiers. All of these classifiers are characteristic of supervised machine learning, meaning that they require some labeled data in order to learn the function

$$f(message, \theta) = \left\{ \begin{array}{cc} \mathrm{C}_{\mathrm{spam}} & \text{if classified as spam} \\ \mathrm{C}_{\mathrm{leg}} & \text{otherwise} \end{array} \right\}$$

where, m is the message to be classified and is a vector of parameters and Cspam and Cleg arerespectively spam and legitimate messages. The task of detecting fake news is similar and almost analogous to the task of spam detection in that both aim to separate examples of legitimate text from examples of illegitimate, ill-intended texts.

2.3.2 STANCE DETECTION

The goal of this contest was to encourage the development of tools that may help human fact checkers identify deliberate misinformation in news stories through the use of machine learning,natural language processing and artificial intelligence. The organizers decided that the first stepin this overarching goal was understanding what other news organizations are saying about thetopic in question. As such, they decided that stage one of their contest would be a stance detection competition. More specifically, the organizers built a dataset of headlines and bodiesof text and challenged competitors to build classifiers that could correctly label the stance of a body text, relative to a given headline, into one of four categories: “agree”, “disagree”,“discusses” or “unrelated.” The top three teams all reached over 80% accuracy on the test set for this task. The top teams model was based on a weighted average between gradient-boosteddecision trees and a deep convolutional neural network.

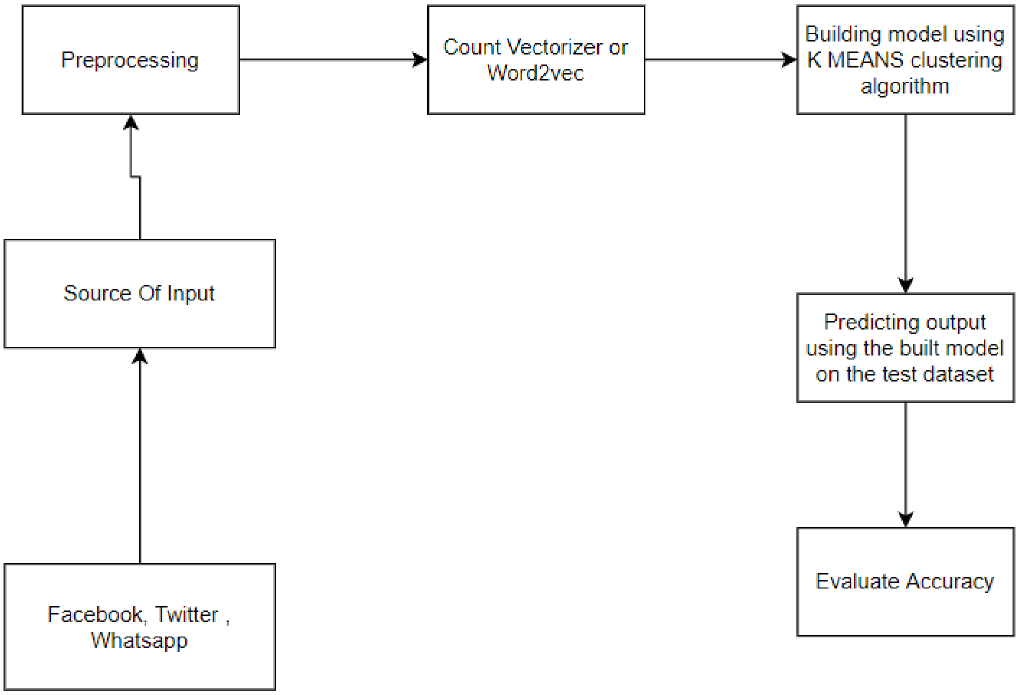
CHAPTER 3

METHODOLOGY

3.1 PROPOSED SYSTEM

The proposed system when subjected to a scenario of a set of news articles , the new articles are categorized as true or fake by the existing data available . This prediction is done byusing the relationship between the words used in the article with one another. The proposed system contains a Word2Vec model for finding the relationship between the words and with theobtained information of the existing relations , the new articles are categorized into fake and real news.

3.2 SYSTEM ARCHITECTURE


Preprocessing
Count Vectorizer or Word2vec
Building model using K MEANS clustering algorithm
Source Of Input
Predicting output using the built model on the test dataset
Facebook, Twitter , Whatsapp
Evaluate Accuracy

Input is collected from various sources such as newspapers , social media andstored in datasets. System will take input from datasets. The datasets undergo preprocessing and the unnecessary information is removed from it and the data types of the columns are changed if required. Jupyter notebook and python libraries are used in theabove step. Count vectorizer technique is used in the initial step. For fake news detection

, we have to train the system using dataset. Before entering to the detection of fake news

, entire dataset is divide into two datasets . 80% is used for training and 20% is used for testing. During training , K-Means algorithm is used to train the model using the train dataset. In testing , the test dataset is given as input and the output is predicted.After the testing time , The predicted output and the actual output are compared using confusion matrix obtained .The confusion matrix gives the information regarding the number of correct and wrong predictions in the case of real and fake news.The accuracy is calculated by the equation No Of Correct Predictions/Total Test Dataset Input Size

3.3 ALGORITHM FOR THE PROPOSED SYSTEM:

Step 1: Start

Step 2: Input is collected from various sources and prepare a dataset.

Step 3: Preprocessing of data is done and dataset is divided into 2 parts training andtesting data.

Step 4: Count vectorization technique is used to convert the train data into numericals.

Step 5: K MEANS clustering algorithm is used to build the predictive model using thetrain data .

Step 6: Confusion matrix is obtained .Step 7: Accuracy is calculated

CHAPTER 4

DATASET

4.1 EXISTING DATASETS FOR THIS SYSTEM:

The lack of manually labeled fake news datasets is certainly a bottleneck for advancing computationally intensive, text-based models that cover a wide array of topics. The dataset forthe fake news challenge does not suit our purpose due to the fact that it contains the ground truthregarding the relationships between texts but not whether or not those texts are actually true orfalse statements. For our purpose, we need a set of news articles that is directly classified into categories of news types (i.e. real vs. fake or real vs parody vs. clickbait vs. propaganda). For more simple and common NLP classification tasks, such as sentiment analysis, there is an abundance of labeled data from a variety of sources including Twitter, Amazon Reviews, and IMDb Reviews. Unfortunately, the same is not true for finding labeled articles of fake and realnews. This presents a challenge to researchers and data scientists who want to explore the topicby implementing supervised machine learning techniques. I have researched the available datasets for sentence-level classification and ways to combine datasets to create full sets with positive and negative examples for document-level classification.

4.2 : PROPOSED DATASET USED:

There exists no dataset of similar quality to the Liar Dataset for document level classification of fake news. As such, I had the option of using the headlines of documents as statements or creating a hybrid dataset of labeled fake and legitimate news articles. Thisshows an informal and exploratory analysis carried out by combining two datasets that individually contain positive and negative fake news examples. Genes trains a model on aspecific subset of both the Kaggle dataset and the data from NYT and the Guardian. In his experiment, the topics involved in training and testing are restricted to U.S News, Politics,Business and World news. However, he does not account for the difference in date range between the two datasets, which likely adds an additional layer of topic bias based on topicsthat are more or less popular during specific periods of time. We have collected data in a manner similar to that of Genes , but more cautious in that we control for more bias in thesources and topics. Because the goal of our project was to find patterns in the language thatare indicative of real or fake news, having source bias would be detrimental to our purpose.Including any source bias in our dataset, i.e. patterns that are specific to NYT, The Guardian, or any of the fake news websites, would allow the model to learn to associate sources with real/fake news labels. Learning to classify sources as fake or real news is an easy problem, but learning to classify specific types of language and language patterns as fake or real news is not. As such, we were very careful to remove as much of 15 the source-specific patterns as possible to force our model to learn something more meaningful and generalizable. We admit that there are certainly instances of fake news in the New York Times and probably instances of real news in the Kaggle dataset because it is based on a list of unreliable websites. However, because these instances are the exception and not therule, we expect that the model will learn from the majority of articles that are consistent with the label of the source. Additionally, we are not trying to train a model to learn facts but rather learn deliveries. To be more clear, the deliveries and reporting mechanisms found in fake news articles within New York Times should still possess characteristics more commonly found in real news, although they will contain fictitious factual information.

4.3: FAKE NEWS SAMPLES:

The system uses a dataset of fake news articles that was gathered by using a tool called the BS detector which essentially has a blacklist of websites that are sources of fakenews. The articles were all published in the 30 days between October, 26 2016 to November25, 2016. While any span of dates would be characterized by the current events of that time,this range of dates is particularly interesting because it spans the time directly before, during, and directly after the 2016 election. The dataset has articles and metadata from 244different websites, which is helpful in the sense that the variety of sources will help the model to not learn a source bias. However, at a first glance of the dataset, you can easily tell that there are still certain obvious reasons that a model could learn specifics of what isincluded in the "body" text in this dataset. For example, there are instances of the author and source in the body text , Also, there are some patterns like including the date that, if not also repeated in the real news dataset, could be learned by the model

CHAPTER 5

CONCEPTS

5.1 PREPROCESSING:

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. Inother words, the features of the data can now be easily interpreted by the algorithm.

In this fake news detection, pre processing is the major thing that should be done . Firstly

, as the data dataset is collected from various sources unnecessary information should beremoved ,converted to lower case , remove punctuation , symbols , stop words.

5.2 STEPS IN TEXT PRE-PROCESSING:

5.2.1 TEXT NORMALIZATION:

Text normalization is a process of transforming text into a single canonical form. Normalizing text before storing or processing it allows for separation of required data fromthe rest so that the system can send consistent data as an input to the other steps of the algorithm.

5.2.2 STOP WORD REMOVAL

5.2.2.1 Stop Word:

A Stop Word is a commonly used word in any natural language such as “a, an , the, for, is,was, which, are, were, from, do, with, and, so, very, that, this, no, yourselves etc. .... ”.

These Stop Words will have a very high frequency and so these should be eliminated whilecalculating the term frequency so that the other important things are given priority. Stop word removal is such a Pre-processing step which removes these stop words and thereby helping in the further steps and also reducing some processing time because the size of thedocument decreases tremendously.

Consider a Sentence

“This is a sample sentence, showing off the stop word removal”.Output after Stop word removal is:

[“sample”, “sentence”, “showing”, “stop”, “word”, “removal”]

Note: Though Stop words refer to the most commonly used words in a particular language,there is no single universal list of stop words, different tools uses different stop words.

5.2.3 STEMMING:

Stemming is a pre-processing step in Text Mining applications as well as a very common requirement of Natural Language processing functions. In fact it is very important in most of the Information Retrieval systems. The main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. Eg: A stemmer for English should identify the strings "cats", "catlike", "catty" as basedon the root "cat".

5.2.3.1 RULES OF SUFFIX STRIPPING STEMMERS:

1.If the word ends in 'ed', remove the 'ed'. 2.If the word ends in 'ing', remove the 'ing'.3.If the word ends in 'ly', remove the 'ly'.

5.2.3.2 RULES OF SUFFIX SUBSTITUTION STEMMERS:

1.If the word ends in ‘ies’ substitute ‘ies’ with ‘y’.

Generally this stemmer is used because of some word like families etc...

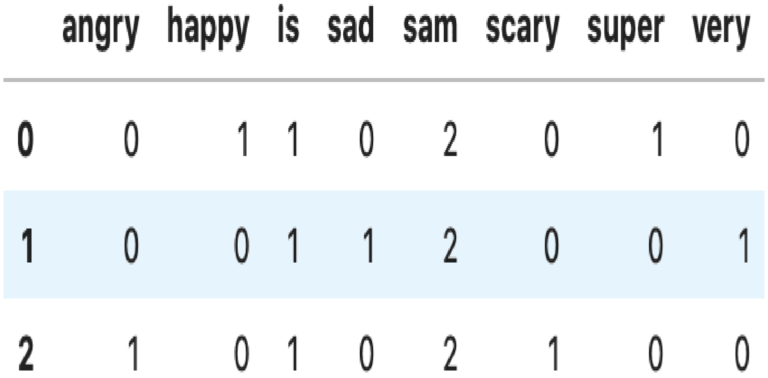
5.3 COUNT VECTORIZER:

CountVectorizer tokenizes(tokenization means breaking down a sentence or paragraph or any text into words) the text along with performing very basic preprocessing like removingthe punctuation marks, converting all the words to lowercase, etc.The vocabulary of known words is formed which is also used for encoding unseen text later.An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of timeseach word appeared in the document.

5.3.1 Input to count vectorizer:

Document having 3 sentencessam sam is super happy

sam sam is very sad sam sam is scary angry**Output:**



	angry	happy	is	sad	sam	scary	super	very
0	0	1	1	0	2	0	1	0
1	0	0	1	1	2	0	0	1
2	1	0	1	0	2	1	0	0

5.4 WORD2VEC MODEL:

WORD2VEC is a class of models that represents a word in a large text corpus as a vector in n-dimensional space bringing similar words closer to each other. One such model is the Skip- Gram model. It can be used to learn word embeddings from large datasets. Embeddings learnedthrough word2vec have proven to be successful on a variety of downstream natural language processing tasks. The context of a word can be represented through a set of skip-gram pairs of (target_word,context_word)where context_word appears in the neighboringcontext of target_word.

5.4.1 WORD2VEC ALGORITHM :

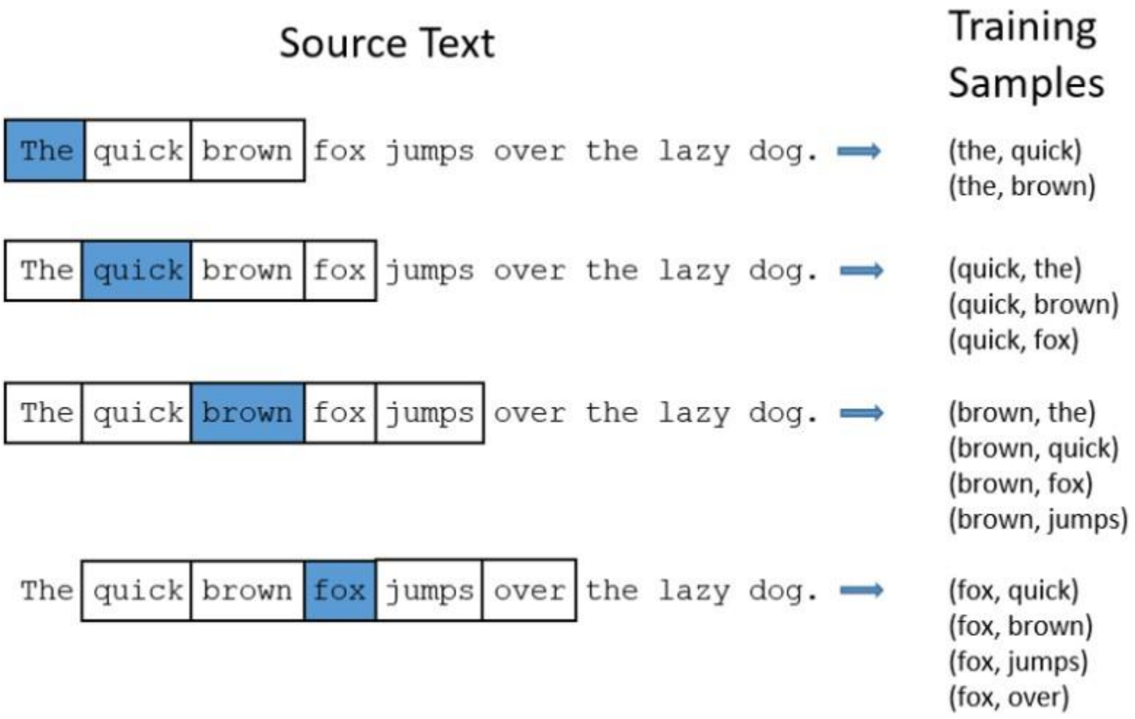
1.Take a sentence as input2.Consider a window size

3. For every word in the sentence 1.Consider current word as context.

2. Other words in the window to the left and right of the word as targets andform (context,target) pair.

3. From the pre defined vocabulary in the tensorflow library , the position of thecontext and target are found and then those values are applied in this formula

4.The output is sent to the sigmoid function to result in the range [-1,1]



$$W_{ij} = \frac{\Sigma_p(x_i-\overline{x_i})(x_j-\overline{x_j})}{\sqrt{\Sigma_p(x_i-\overline{x_i})^2}\sqrt{\Sigma_p(x_j-\overline{x_j})^2}}$$

5.5 K-MEANS ALGORITHM :

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.A cluster refers to a collection of datapoints aggregated together because of certain similarities.You’ll define a target number k, whichrefers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, whilekeeping the centroids as small as possible.The ‘means’ in the K-means refers to averaging of thedata; that is, finding the centroid.To process the learning data, the K-means algorithm in machine learning starts with a first group of randomly selected centroids, which are used as the beginningpoints for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when :

- The centroids have stabilized — there is no change in their values because the clusteringhas been successful.

5.6 EVALUATION MEASURES:

Whenever we build Machine Learning models, we need some form of metric to measure the goodness of the model. Bear in mind that the "goodness" of the model could have multiple interpretations, but generally when we speak of it in a Machine Learning context we are talkingof the measure of a model's performance on new instances that weren't a part of the training data.

Determining whether the model being used for a specific task is successful depends on 2 key factors:

1. Whether the evaluation metric we have selected is the correct one for our problem
2. If we are following the correct evaluation process

In this article, I will focus only on the first factor — Selecting the correct evaluation metric.

5.6.1 DIFFERENT TYPES OF EVALUATION METRICS

The evaluation metric we decide to use depends on the type of NLP task that we are doing. To further add, the stage the project is at also affects the evaluation metric we are using. For instance,during the model building and deployment phase, we'd more often than not use a different evaluation metric to when the model is in production. In the first 2 scenarios, ML metrics wouldsuffice but in production, we care about business impact, therefore we'd rather use business metrics to measure the goodness of our model.

With that being said, we could categorize evaluation metrics into 2 buckets.

- Intrinsic Evaluation — Focuses on intermediary objectives (i.e. the performance ofan NLP component on a defined subtask)
- Extrinsic Evaluation — Focuses on the performance of the final objective (i.e. theperformance of the component on the complete application)

Stakeholders typically care about extrinsic evaluation since they'd want to know how good themodel is at solving the business problem at hand. However, it's still important to have intrinsicevaluation metrics in order for the AI team to measure how they are doing. We will be focusingmore on intrinsic metrics for the remainder of this article.

5.6.2 DEFINING THE METRICS

Some common intrinsic metrics to evaluate NLP systems are as follows:

5.6.2.1 ACCURACY

Whenever the accuracy metric is used, we aim to learn the closeness of a measured value to a known value. It’s therefore typically used in instances where the output variable is categorical ordiscrete — Namely a classification task.

5.6.2.2 PRECISION

In instances where we are concerned with how exact the model's predictions are we would use Precision. The precision metric would inform us of the number of labels that are actually labeledas positive in correspondence to the instances that the classifier labeled as positive.

5.6.2.3 RECALL

Recall measures how well the model can recall the positive class (i.e. the number of positive labels that the model identified as positive

CHAPTER 6

EXPERIMENT ANALYSIS

6.1 SYSTEM CONFIGURATION

This project can run on commodity hardware. We ran entire project on an Intel I5 processor with 8 GB Ram, 2 GB Nvidia Graphic Processor, It also has 2 cores which runs at 1.7GHz, 2.1 GHz respectively. First part of the is training phase which takes 10-15 mins of time and the second part is testing part which only takes few

seconds to make predictions and calculate accuracy.

6.1.1 HARDWARE REQUIREMENTS:

- RAM: 4 GB
- Storage: 500 GB
- CPU: 2 GHz or faster
- Architecture: 32-bit or 64-bit

6.1.2 SOFTWARE REQUIREMENTS

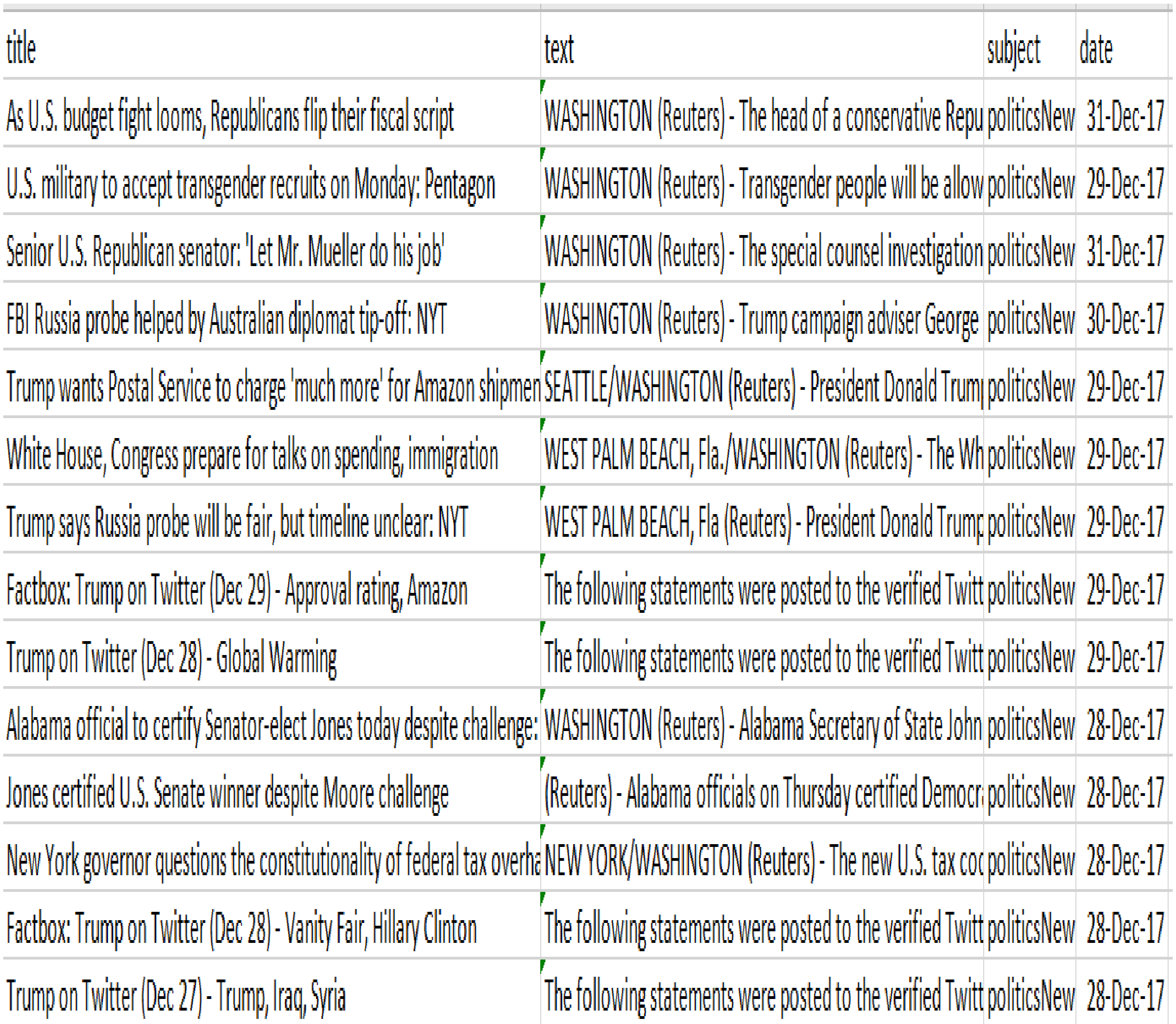
- Python 3.5 in Google Colab is used for data pre-processing, model training and prediction.
- Operating System: windows 7 and above or Linux based OS or MAC OS.

6.2 Sample input

The dataset contains 4 columns

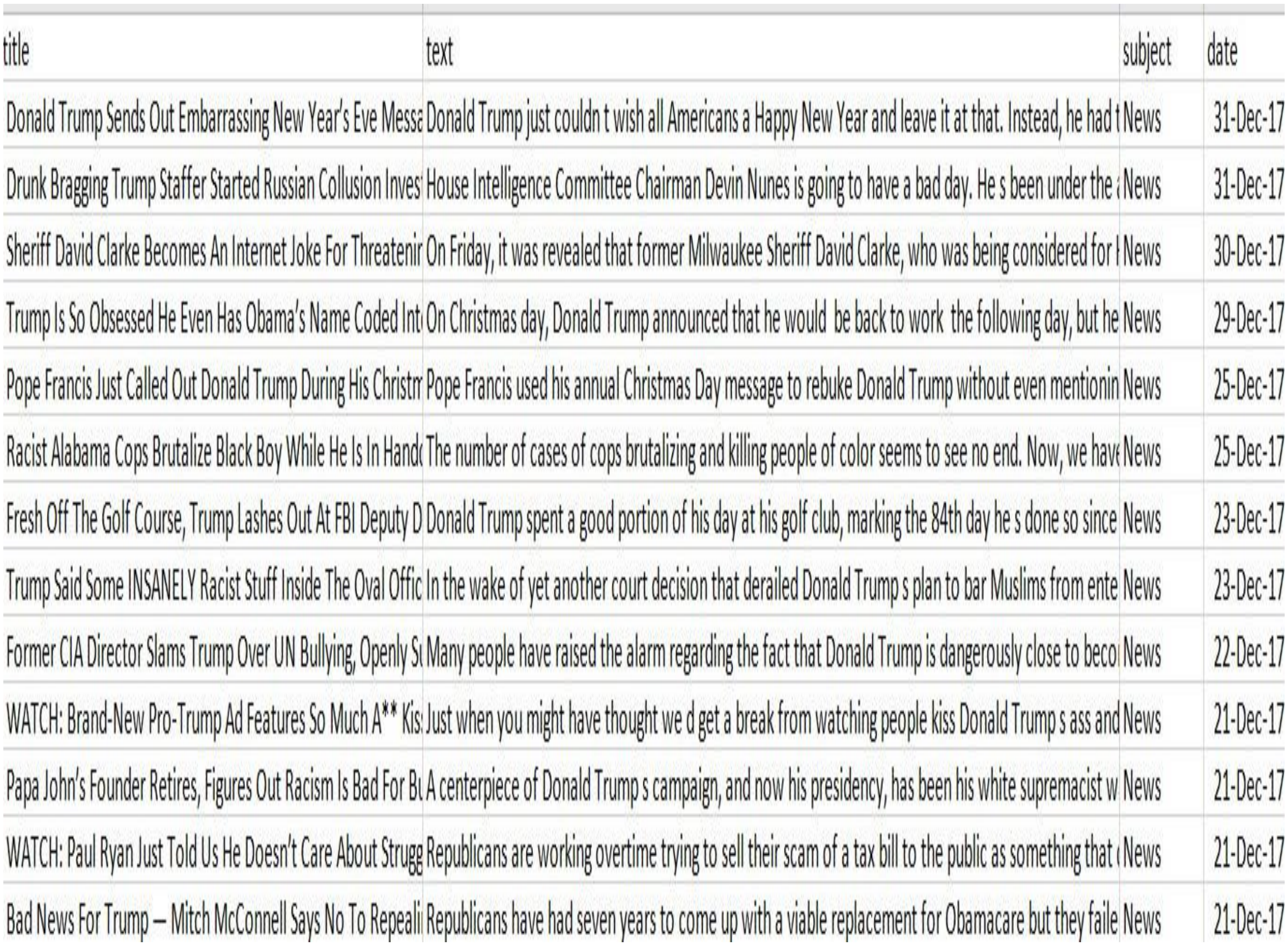
1. Title
2. Text
3. Subject
4. Date

True.csv



title	text	subject	date
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Repu	politicsNew	31-Dec-17
U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allow	politicsNew	29-Dec-17
Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	WASHINGTON (Reuters) - The special counsel investigation	politicsNew	31-Dec-17
FBI Russia probe helped by Australian diplomat tip-off: NYT	WASHINGTON (Reuters) - Trump campaign adviser George	politicsNew	30-Dec-17
Trump wants Postal Service to charge 'much more' for Amazon shipmen	SEATTLE/WASHINGTON (Reuters) - President Donald Trum	politicsNew	29-Dec-17
White House, Congress prepare for talks on spending, immigration	WEST PALM BEACH, Fla./WASHINGTON (Reuters) - The Wh	politicsNew	29-Dec-17
Trump says Russia probe will be fair, but timeline unclear: NYT	WEST PALM BEACH, Fla (Reuters) - President Donald Trump	politicsNew	29-Dec-17
Factbox: Trump on Twitter (Dec 29) - Approval rating, Amazon	The following statements were posted to the verified Twitt	politicsNew	29-Dec-17
Trump on Twitter (Dec 28) - Global Warming	The following statements were posted to the verified Twitt	politicsNew	29-Dec-17
Alabama official to certify Senator-elect Jones today despite challenge:	WASHINGTON (Reuters) - Alabama Secretary of State John	politicsNew	28-Dec-17
Jones certified U.S. Senate winner despite Moore challenge	(Reuters) - Alabama officials on Thursday certified Democr	politicsNew	28-Dec-17
New York governor questions the constitutionality of federal tax overha	NEW YORK/WASHINGTON (Reuters) - The new U.S. tax co	politicsNew	28-Dec-17
Factbox: Trump on Twitter (Dec 28) - Vanity Fair, Hillary Clinton	The following statements were posted to the verified Twitt	politicsNew	28-Dec-17
Trump on Twitter (Dec 27) - Trump, Iraq, Syria	The following statements were posted to the verified Twitt	politicsNew	28-Dec-17

Fake.csv



title	text	subject	date
Donald Trump Sends Out Embarrassing New Year's Eve Messa	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had t	News	31-Dec-17
Drunk Bragging Trump Staffer Started Russian Collusion Inves	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the a	News	31-Dec-17
Sheriff David Clarke Becomes An Internet Joke For Threatenir	On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for I	News	30-Dec-17
Trump Is So Obsessed He Even Has Obama's Name Coded Int	On Christmas day, Donald Trump announced that he would be back to work the following day, but he	News	29-Dec-17
Pope Francis Just Called Out Donald Trump During His Christm	Pope Francis used his annual Christmas Day message to rebuke Donald Trump without even mentionin	News	25-Dec-17
Racist Alabama Cops Brutalize Black Boy While He Is In Hand	The number of cases of cops brutalizing and killing people of color seems to see no end. Now, we have	News	25-Dec-17
Fresh Off The Golf Course, Trump Lashes Out At FBI Deputy D	Donald Trump spent a good portion of his day at his golf club, marking the 84th day he s done so since	News	23-Dec-17
Trump Said Some INSANELY Racist Stuff Inside The Oval Offic	In the wake of yet another court decision that derailed Donald Trump s plan to bar Muslims from ente	News	23-Dec-17
Former CIA Director Slams Trump Over UN Bullying, Openly S	Many people have raised the alarm regarding the fact that Donald Trump is dangerously close to beco	News	22-Dec-17
WATCH: Brand-New Pro-Trump Ad Features So Much A** Kis	Just when you might have thought we d get a break from watching people kiss Donald Trump s ass and	News	21-Dec-17
Papa John's Founder Retires, Figures Out Racism Is Bad For Bu	A centerpiece of Donald Trump s campaign, and now his presidency, has been his white supremacist w	News	21-Dec-17
WATCH: Paul Ryan Just Told Us He Doesn't Care About Strugg	Republicans are working overtime trying to sell their scam of a tax bill to the public as something that (	News	21-Dec-17
Bad News For Trump – Mitch McConnell Says No To Repeali	Republicans have had seven years to come up with a viable replacement for Obamacare but they faile	News	21-Dec-17

6.3 SAMPLE CODE:

6.3.1 IMPORTING THE LIBRARIES:

```
import numpy as npimport pandas as pd

import matplotlib.pyplot as pltimport seaborn as sns
sns.set() import stringimport re

from gensim.parsing.preprocessing import preprocess_string, strip_tag s,
strip_punctuation, strip_multiple_whitespaces, strip_numeric, remov
e_stopwords, strip_short
from gensim.models import Word2Vec

from sklearn import cluster from sklearn import metrics
from sklearn.decomposition import PCAfrom sklearn.manifold import TSNE
```

6.3.1 READING THE DATASETS:

```
fake = pd.read_csv('/content/drive/MyDrive/Fake.csv')true =
pd.read_csv('/content/drive/MyDrive/True.csv')
```

6.3.2 FIND THE NULL VALUES:

```
print(fake.isnull().sum())print('************') print(true.isnull().sum())
```

6.3.3 FILL THE NULL VALUES:

```
true=true.fillna(' ') fake=fake.fillna(' ')
```

6.3.4 REMOVE UNNECESSARY DATA:

```
  cleansed_data = [] for data in true.text:
if "@realDonaldTrump : - " in data:
  cleansed_data.append(data.split("@realDonaldTrump : - ")[1])
```

```
elif "(Reuters) -" in data: cleansed_data.append(data.split("(Reuters) - ")[1])
 else:
 cleansed_data.append(data)

 true["text"] = cleansed_datatrue.head(10)
```

6.3.5 CLUB TEXT AND TITLE:

```
fake['Sentences'] = fake['title'] + ' ' + fake['text']true['Sentences'] = true['title']
+ ' ' + true['text']
```

6.3.6 ASSIGN LABELS FOR THE TEXT:

```
fake['Label'] = 0
true['Label'] = 1
```

6.3.6 CONCATINATING TWO DATASETS:

```
final_data = pd.concat([fake, true])

final_data = final_data.sample(frac=1).reset_index(drop=True)
```

```
        final_data = final_data.drop(['title', 'text', 'subject', 'date'], axis = 1)
```

6.3.7 CATEGORIZING WORDS TO REAL AND FAKE:

```
        real_words = ''fake_words = ''

    for val in final_data[final_data['Label']==1].Sentences:# split the value

        tokens = val.split()

        # Converts each token into lowercasefor i in range(len(tokens)):

            tokens[i] = tokens[i].lower() real_words += " ".join(tokens)+" "

    for val in final_data[final_data['Label']==0].Sentences:# split the value

        tokens = val.split()

        # Converts each token into lowercasefor i in range(len(tokens)):
```

```
    tokens[i] = tokens[i].lower() fake_words += " ".join(tokens)+" "
```

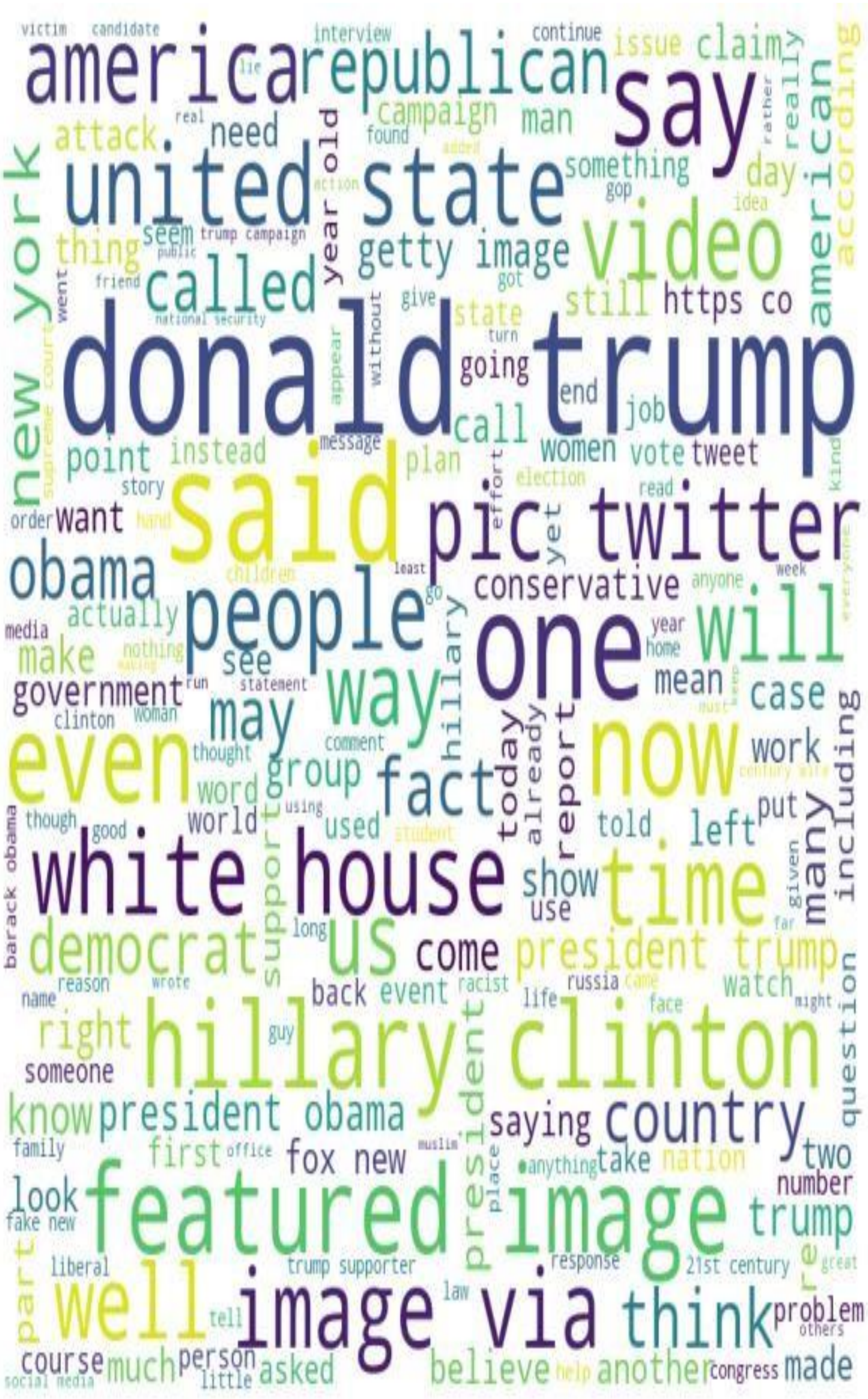
6.3.7 VISUALIZE REAL WORDS:

```
from wordcloud import WordCloud, STOPWORDSfrom nltk.corpus import

stopwords

stopwords = set(STOPWORDS)

ordcloud = WordCloud(width = 800, height = 800,background_color ='white',
stopwords = stopwords,
min_font_size = 10).generate(real_words)


# plot the WordCloud image plt.figure(figsize = (8, 8), facecolor = None)

plt.imshow(wordcloud)

plt.axis("off") plt.tight_layout(pad = 0)plt.show()
```

despite
problem
talk
result
part
working
bill
area
year old
decision
many
deal
meeting
without
last year
us
news conference
believe
city
vote
way
vice president
comment
view
officials said
come
president elect
white house
wednesday
week
turkey
others
germany
republican presidential
national security
use
may
islamic
run
state
thursday
concern
even
seen
countries
hope
lead
people
protest
added
britain
used
debate
much
far
likely
said
country
want
home
order
time
mean
measure
friday
remain
move
head
march
show
following
effort
military
hillary clinton
call
saying
say
plan
report
going
support
one
change
case
said trump
state
sunday
said wednesday
russia
presidential election
trump campaign
new york
last week
asked
official said
government
secretary
republican senator
voter
called
need
told
reporter
united nation
general
attorney
although
rule
iran
expected
issue
number
said monday
will
put
european union
region
according
last month
united state
washington
give
saturday
american
state department
eu
might
make
around
include
said statement
told reuters
back
house representative
left
proposal
human right
north korean
south
president
donald
event
end
said friday
attack
help
program
trump said
group
first
economy
including
monday
among
said tuesday
congress
company
position
democrat
several
south korea
trump administration
still
long
donald trump
think
made
face
new
yet
day
possible
action
see
side
saudi arabia
prime minister
barack obama
now
july
said thursday
companie
president barack
work
another
political
power
obama administration
foreign minister
north korea
law
clinton
two
given
tuesday
well
set
allow
presidential candidate
already
lawmaker
china
member
supreme court
job

6.3.8 VISUALIZE FAKE WORDS:

```
ordcloud = WordCloud(width = 800, height = 800,background_color ='white',
        stopwords = stopwords,
        min_font_size = 10).generate(fake_words)


        # plot the WordCloud image plt.figure(figsize = (8, 8), facecolor = None)

        plt.imshow(wordcloud)

        plt.axis("off") plt.tight_layout(pad = 0)



        plt.show()
```


victim
candidate
interview
continue
america
republican
say
issue
claim
really
according
lie
real
attack
need
campaign
man
found
added
rather
american
united
state
something
day
year old
action
gop
idea
thing
seem
trump campaign
public
getty image
video
friend
went
got
called
give
state
still
https co
national security
new york
donald
trump
turn
without
appear
going
end
job
supreme court
call
message
point
instead
plan
women
vote
tweet
story
effort
election
read
kind
order
want
hand
said
pic
twitter
everyone
obama
children
least
yet
conservative
anyone
week
media
actually
people
go
hillary
one
will
year
home
make
nothing
making
see
way
mean
government
run
statement
keep
case
clinton
woman
must
may
already
report
even
thought
comment
today
now
work
century wife
word
group
fact
including
many
though
using
put
barack obama
good
world
used
student
told
left
white house
support
show
time
given
use
long
far
us
come
democrat
president trump
reason
wrote
back
event
racist
russia
came
watch
name
life
face
might
question
right
hillary clinton
guy
president
someone
know
president obama
saying
country
family
first
office
muslim
fox new
place
anything
take
nation
two
number
look
featured image
trump
fake new
part
re
liberal
trump supporter
response
21st century
great
law
image
via
think
problem
tell
others
course
much
person
asked
believe
help
another
congress
made
social media
little

6.3.9 PRE PROCESSING THE TEXT:

To remove urls

```
def remove_URL(s):

regex = re.compile(r'https?://\S+|www\.\S+|bit\.ly\S+')return regex.sub(r'',s)
```

1.To convert text to lower case - x.lower() 2.Remove unneseccary spaces at the end - strip_tags3.To remove url – Above function

4. To remove punctuation – strip_punctuation

5. To remove multiple white spaces in the sentence betweenwords – strip_multiple_whitespaces

6. To remove numbers – strip_numeric

7. To remove stopwords – remove_stopword

```
CUSTOM_FILTERS = [lambda x: x.lower(), strip_tags, remov e_URL,
strip_punctuation, strip_multiple_whitespaces, strip_n umeric,
remove_stopwords, strip_short]
```

```
processed_data = [] processed_labels = []

for index, row in final_data.iterrows():

    words_broken_up = preprocess_string(row['Sentences'], CU

STOM_FILTERS)

if len(words_broken_up) > 0: processed_data.append(words_broken_up)

processed_labels.append(row['Label'])


print(len(processed_data)) # train=35912

# test=8977
```

Output of one article after pre processing:

```
['bikers', 'trump', 'travel', 'future', 'rallies', '“provide', 'outside', 'security”',
'paid', 'soros', 'thugs', 'hillary', 'bernie', 'sanders', 'americans', 'know', 'come',
'anarchists', 'whiny', 'petulant', 'college', 'students', 'better', 'angry', 'blm','protesters',
'meet', 'group', 'care', 'feelings', 'political', 'correctness', 'large',
```

'percentage', 'bikers', 'belong', 'groups', 'veterans', 'fought', 'nation', 'step', 'aside', 'allow', 'billionaire', 'communist', 'supports', 'woman', 'investigation', 'fbi', 'left', 'brothers', 'die', 'benghazi', 'away', 'right', 'americans', 'attend', 'political', 'rally', 'candidate', 'support', 'military', 'tradition', 'sorts', 'running', 'decades', 'wake', 'world', 'war', 'generation', 'troops', 'returned', 'home',

'combat', 'veterans', 'country', 'certain', 'pleasure', 'purpose', 'newly', 'evolved',

'piece', 'gear', 'friendly', 'downrange', 'motorcycle', 'new', 'motorcycle', 'clubs', 'sprang', 'filling', 'void', 'camaraderie', 'brotherhood', 'mention', 'adrenaline', 'adventure', 'craving', 'end', 'military', 'service', 'called', 'outlaws', 'criminals', 'refused', 'boxed', 'rules', 'regulations', 'fledgling', 'american', 'motorcycle', 'association', 'combat', 'motorcycle', 'outlaws', 'come', 'says', 'charles', 'davis',

'writes', 'aging', 'rebel', 'biker', 'news', 'blog', 'los', 'angeles', 'clubs', 'like','boozefighters', 'outlaws', 'invented', 'transformed', 'veterans', 'cheap', 'army', 'surplus', 'bikes', 'club', 'particular', 'drew', 'inspiration', 'pursuit', 'squadron', 'flying', 'tigers', 'american', 'volunteers', 'flew', 'combat', 'missions', 'japanese','china', 'squadron', 'better', 'known', 'fliers', 'hells', 'angels', 'waves',

'motorcycle', 'club', 'membership', 'davis', 'says', 'second', 'surge', 'corps',

'artilleryman', 'arose', 'wake', 'vietnam', 'like', 'war', 'fighters', 'returning',

'home', 'largely', 'hostile', 'nation', 'family', 'bikers', 'clubs', 'mongols', 'devils',

'disciples', 'named', 'george', 'bernard', 'shaw', 'play', 'revolutionary', 'war',

'patriot', 'ethan', 'allen', 'davis', 'says', 'bandidos', 'got', 'start', 'largely', 'fueled','returning', 'veterans', 'new', 'generation', 'currently', 'serving', 'troops','veterans', 'pouring', 'old', 'clubs', 'starting', 'groups', 'military', 'timesmeanwhile', 'donald', 'trump', 'ending', 'vacation', 'rally', 'critical', 'state', 'wisconsin', 'tomorrow', 'event', 'sold', 'violent', 'protest', 'organized', 'cause',

'mayhem', 'havoc', 'arizona', 'illinois', 'quote', 'box', 'center', 'trump', 'patriots',

'facebook', 'page', 'patriotic', 'bikers', 'united', 'states', 'planning', 'future',

```
'trump', 'rallies', 'sure', 'paid', 'agitator', 'protesters', 'away', 'trump', 'right',
'speak', 'interfere', 'rights', 'trump', 'supporters', 'safely', 'attend', 'shall', 'silenced', 'paid',
'protestors', 'planning', 'causing', 'chaos', 'violence', 'anarchy', 'riots', 'trump', 'rallies',
'private', 'paid', 'events', 'private', 'property', 'trump', 'secret', 'service', 'protection', 'want',
'peacefully', 'assemble', 'street', 'trumps', 'rallies', 'protest', 'amendment', 'right', 'publicly',
'plan', 'incite', 'organize','events', 'paid', 'agitators', 'disrupting', 'civil', 'rights', 'attending',
'private','event', 'likely', 'end', 'bad', 'despite', 'medias', 'attempt', 'cheerleaders', 'quote',
'box', 'center', 'janesville', 'nichole', 'mittness', 'thought', 'people', 'respond',
'facebook', 'page', 'inviting', 'protest', 'donald', 'trump', 'janesville', 'appearance', 'midday',
'saturday', 'pledged', 'mittness', 'figured', 'meant', 'tuesday', 'overwhelming', 'anticipating',
'kind', 'response', 'mittness', 'said', 'mittness', 'working', 'peaceful', 'protest', 'interfere',
'trump', 'event', 'janesville','police', 'preparing', 'possibility', 'janesville', 'police', 'chief',
'dave', 'moore', 'said', 'friday', 'know', 'officers', 'assisgned', 'department', 'reached',
'police',
'agencies', 'rock', 'county', 'including', 'sheriff', 'office', 'state', 'patrol', 'dnr',
'dane', 'county', 'sheriff', 'office', 'joint', 'beloit', 'janesville', 'rock', 'county',
'sheriff', 'mobile', 'field', 'force', 'specializes', 'crowd', 'control', 'moore', 'said',
'moore', 'noted', 'janesville', 'conference', 'center', 'holds', 'said', 'expects',
'substantial', 'number', 'people', 'outside', 'trump', 'event', 'scheduled', 'local', 'protest',
'slated', 'begin', 'police', 'respect', 'constitutional', 'right', 'freedom', 'speech', 'degree',
'possible', 'intend', 'allow', 'citizens', 'voice', 'opinions',
'require', 'peaceful', 'safe', 'manner', 'moore', 'said', 'inside', 'janesville',
'conference', 'center', 'holiday', 'inn', 'express', 'different', 'story', 'moore', 'said',
'holiday', 'inn', 'trump', 'people', 'secret', 'service', 'want', 'disrupters', 'removed','private',
'property', 'right', 'moore', 'said', 'prntly']
```

6.3.10 CALCULATING THE DIVISION POINT:

```
import math

#for training 80 percent of data is used

trainlen=math.ceil((4*len(processed_data))/5) print(trainlen)


#for testing 20 percent of data is used


testlen=len(processed_data)-trainlen print(testlen)
```

6.3.11 DIVIDING THE DATASETS:

```
train=processed_data[:trainlen]

test=processed_data[trainlen:] out=final_data.Sentences[trainlen:]
```

```
    print(len(test)) print(out[35912])print(test[0])
```

6.3.12 BUILDING A AWORD2VEC MODEL:

```
    # Word2Vec model trained on processed datamodel = Word2Vec(train,

    min_count=1)
```

6.3.13 FINDING THE SENTENCE VECTOR:

```
def ReturnVector(x):try:

        return model[x]except:

    return np.zeros(100)


    def Sentence_Vector(sentence):

      word_vectors = list(map(lambda x: ReturnVector(x), sentence))
```

```
return np.average(word_vectors, axis=0).tolist()


X = []

                                                  for data_x in test:# print(data_x)

  X.append(Sentence_Vector(data_x)) print(test[0])

X_np = np.array(X)X_np.shape
```

6.3.15 K MEANS ON TEST DATASET:

```
kmeans = cluster.KMeans(n_clusters=2, verbose=0)clustered =


kmeans.fit_predict(X_np)
```

Output of k means:

```
Initialization completestart iteration
done sorting end inner loop
```

```
Iteration 0, inertia 82028.86010166007start iteration
done sorting end inner loop
Iteration 1, inertia 80453.14173407092start iteration
done sorting end inner loop
Iteration 2, inertia 80352.45862835528start iteration
done sorting end inner loop
Iteration 3, inertia 80334.49187040864start iteration
done sorting end inner loop
Iteration 4, inertia 80332.45213793483start iteration
done sorting end inner loop
Iteration 5, inertia 80331.83128468881start iteration
done sorting end inner loop
Iteration 6, inertia 80331.73928403885start iteration
done sorting end inner loop
Iteration 7, inertia 80331.72886884258
center shift 6.268061e-07 within tolerance 1.071682e-05Initialization complete
```

```
start iteration done sorting end inner loop
Iteration 0, inertia 86697.44089445495start iteration
done sorting end inner loop
Iteration 1, inertia 82263.31458487756start iteration
done sorting end inner loop
Iteration 2, inertia 80931.57956219556start iteration
done sorting end inner loop
Iteration 3, inertia 80544.90508632803start iteration
done sorting end inner loop
Iteration 4, inertia 80410.68749230007start iteration
done sorting end inner loop
Iteration 5, inertia 80363.20584129189start iteration
done sorting end inner loop
Iteration 6, inertia 80343.41052152864start iteration
done sorting end inner loop
```

```
Iteration 7, inertia 80336.80275546179start iteration
done sorting end inner loop
Iteration 8, inertia 80334.25214796716start iteration
done sorting end inner loop
Iteration 9, inertia 80332.67364553266start iteration
done sorting end inner loop
Iteration 10, inertia 80332.29084377097start iteration
done sorting end inner loop
Iteration 11, inertia 80331.86731549243start iteration
done sorting end inner loop
Iteration 12, inertia 80331.72785439485start iteration
done sorting end inner loop
Iteration 13, inertia 80331.72785439485
center shift 0.000000e+00 within tolerance 1.071682e-05Initialization complete
start iteration done sorting end inner loop
Iteration 0, inertia 84467.10395991792
```

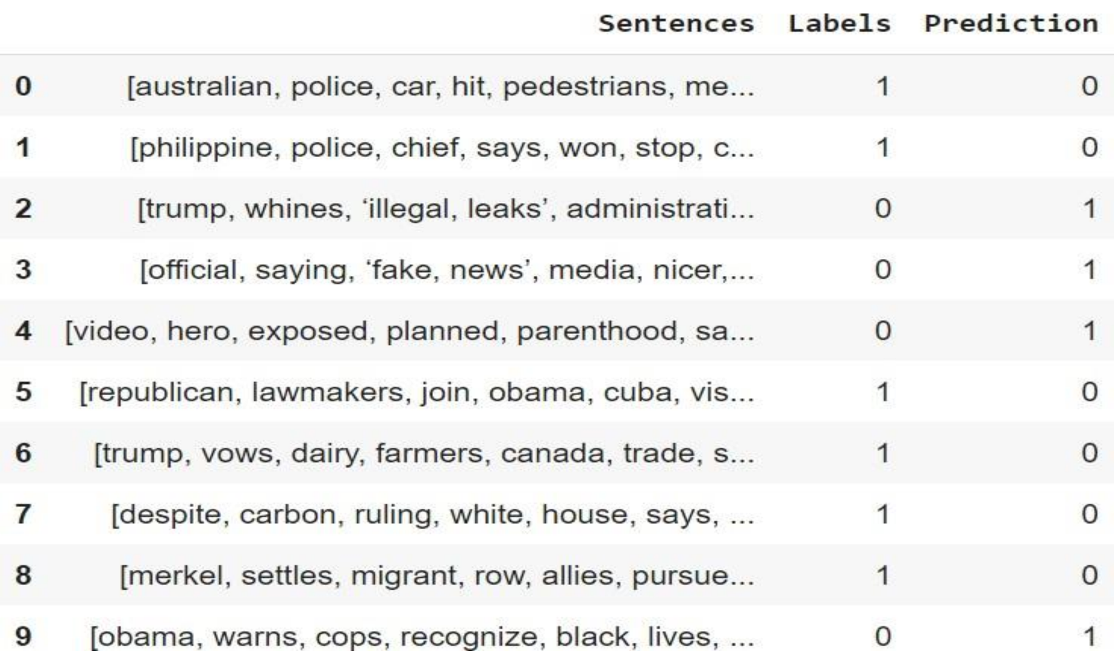
```
start iteration done sorting end inner loop
Iteration 1, inertia 81082.01805089386start iteration
done sorting end inner loop
Iteration 7, inertia 80336.80275546179start iteration
done sorting end inner loop
Iteration 8, inertia 80334.25214796716start iteration
done sorting end inner loop
Iteration 9, inertia 80332.67364553266start iteration
done sorting end inner loop
Iteration 10, inertia 80332.29084377097start iteration
done sorting end inner loop
Iteration 11, inertia 80331.86731549243start iteration
done sorting end inner loop
Iteration 12, inertia 80331.72785439485start iteration
done sorting end inner loop
```

```
Iteration 13, inertia 80331.72785439485
center shift 0.000000e+00 within tolerance 1.071682e-05Initialization complete
start iteration done sorting end inner loop
Iteration 0, inertia 84467.10395991792
Iteration 7, inertia 80336.80275546179start iteration
done sorting end inner loop
Iteration 8, inertia 80334.25214796716start iteration
done sorting end inner loop
Iteration 9, inertia 80332.67364553266start iteration
done sorting end inner loop
Iteration 10, inertia 80332.29084377097start iteration
done sorting end inner loop
Iteration 11, inertia 80331.86731549243start iteration
done sorting end inner loop
Iteration 12, inertia 80331.72785439485start iteration
done sorting end inner loop
```

```
Iteration 13, inertia 80331.72785439485
center shift 0.000000e+00 within tolerance 1.071682e-05Initialization complete
start iteration done sorting end inner loop
Iteration 0, inertia 84467.10395991792
```

6.3.15 PREDICTING THE OUTPUT:

```
testing_df = {'Sentences': test, 'Labels': processed_labels[35912:], 'Prediction':
clustered}
testing_df = pd.DataFrame(data=testing_df)testing_df.head(10)
```



	Sentences	Labels	Prediction
0	[australian, police, car, hit, pedestrians, me...	1	0
1	[philippine, police, chief, says, won, stop, c...	1	0
2	[trump, whines, ‘illegal, leaks’, administrati...	0	1
3	[official, saying, ‘fake, news’, media, nicer,...	0	1
4	[video, hero, exposed, planned, parenthood, sa...	0	1
5	[republican, lawmakers, join, obama, cuba, vis...	1	0
6	[trump, vows, dairy, farmers, canada, trade, s...	1	0
7	[despite, carbon, ruling, white, house, says, ...	1	0
8	[merkel, settles, migrant, row, allies, pursue...	1	0
9	[obama, warns, cops, recognize, black, lives, ...	0	1

6.3.16 COMPARING ORIGINAL TO PREDICTED OUTCOMES:

```
trueneg=truepos=falseneg=falsepos=0for index, row in testing_df.iterrows():
  if row['Labels'] == row['Prediction']==0:trueneg+=1
    if row['Labels'] == row['Prediction']==1:truepos+=1
    if row['Labels'] ==1 and row['Prediction']==0:falseneg+=1
    if row['Labels'] ==0 and row['Prediction']==1:falsepos+=1
print("Correctly clustered news: " + str(((truepos+trueneg)*100)/(trueneg+truepos+falseneg+falsepos)) + "%")
```

Output:

```
Correctly clustered news: 13.0555864988303444%
```

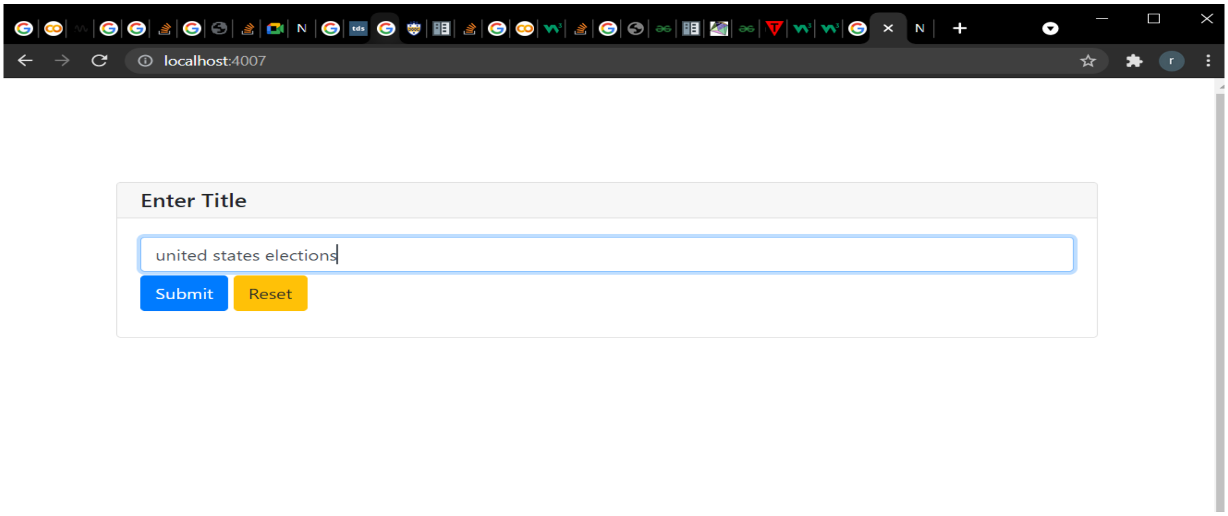
6.3.16 OUTPUT:

```
CONFUSION MATRIX

print(trueneg,falsepos,sep="                    ")print(falseneg,truepos,sep="

                                                ")
```

```
615          4040
3765          557
```

CHAPTER 8 USER INPUT

Enter the title of the article

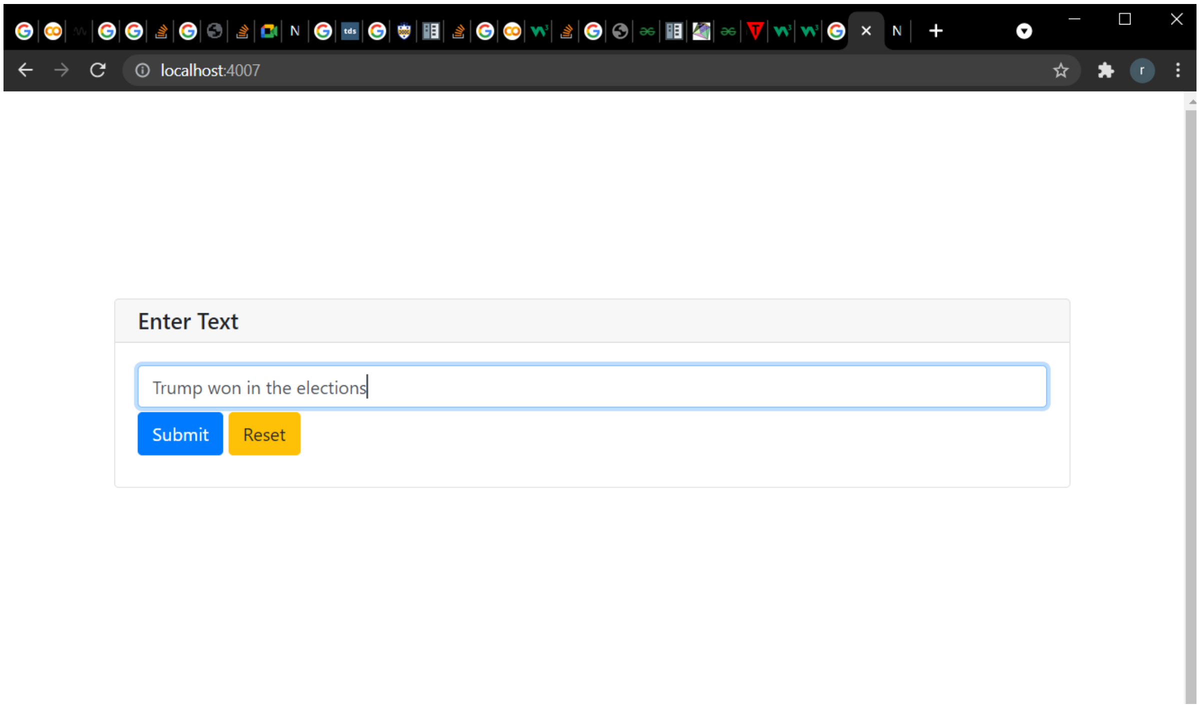


localhost:4007

Enter Title

united states elections

Enter the text of the article

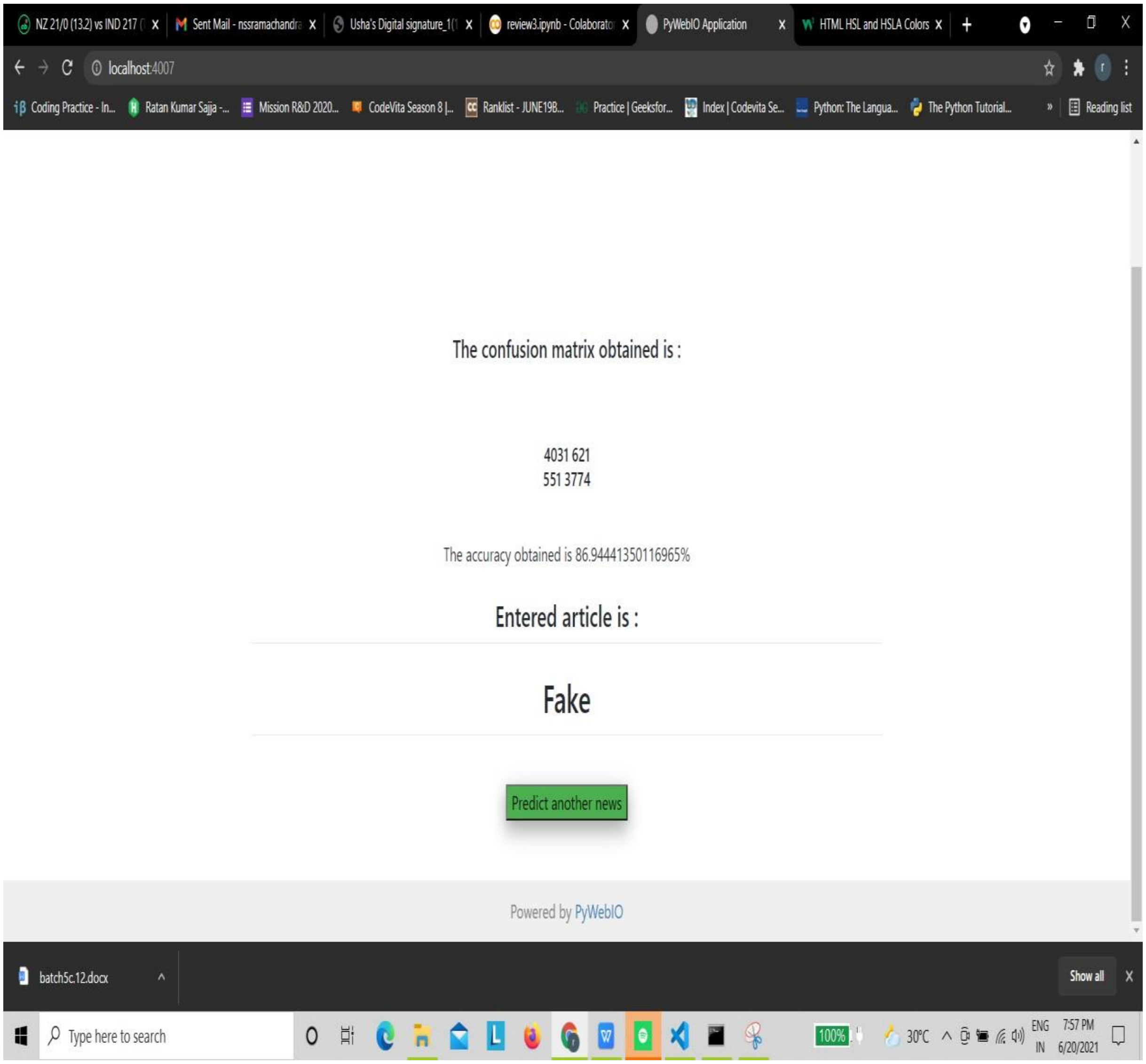


localhost:4007

Enter Text

Trump won in the elections

OUTPUT



```
The confusion matrix obtained is :

4031 621
551 3774

The accuracy obtained is 86.94441350116965%
```

Entered article is :

Fake

CHAPTER 7

CONCLUSION:

In this project, we are predicting whether an article is a real or fake article based on the relationship between the words. We have used the 2016 US president election datasets for creation of this system. We used Word2Vec model for building model and K -Means for the prediction and obtained an accuracy of 87%.

APPENDIX

List of stop words

a	about	above	after	again
against	all	am	an	and
any	are	aren't	as	at
be	because	been	before	being
below	between	both	but	by
can't	cannot	could	couldn't	did
didn't	do	does	doesn't	doing
don't	down	during	each	few
for	From	further	had	hadn't
has	hasn't	have	haven't	having
he	he'd	he'll	he's	her
here	here's	hers	herself	him
himself	his	how	how's	i
i'd	i'll	i'm	i've	if
in	into	is	isn't	It
it's	its	itself	Let's	me
more	most	mustn't	my	myself
no	nor	not	of	off
on	once	only	or	other

ought	our	ours	ourselves	out
over	own	same	shan't	She
she'd	she'll	she's	should	shouldn't
so	some	such	than	that
that's	the	their	theirs	Them
themselves	then	there	there's	These
they	they'd	they'll	they're	they've
this	those	through	To	Too
under	until	up	very	was
wasn't	we	we'd	we'll	we're
we've	were	weren't	what	what's
when	when's	where	where's	which
while	who	who's	whom	why
why's	with	won't	would	wouldn't
you	you'd	you'll	you're	you've
your	yours	Yourself	yourselves	----------

REFERENCES

Datasets: True.csv , Fake.csv

1. International journal of recent technology and engineering (IJRTE) ISSN: 2277-3878,volume-7, issue-6, march 2019

2. Building a fake news classifier using natural language processing BY NATHAN (https://towardsdatascience.com/building-a-fake-news-classifier-using-natural- language-processing-83d911b237e1)

3. Fake news detector: NLP project by ishant juyal (https://levelup.gitconnected.com/fake-news-detector-nlp-project-9d67e0177075)

4. Shloka Gilda,“Evaluating Machine Learning Algorithms for Fake News Detection” ,2017 IEEE 15th Student Conference on Research and Development (SCOReD).

5. Mykhailo Granik, Volodymyr Mesyura, “Fake News Detection Using Naive BayesClassifier”, 2017 IEEEFirst Ukraine Conference on Electrical and Computer Engineering (UKRCON).

6. Gravanis, G., et al., Behind the cues: A benchmarking study for fake news detection.Expert Systems with Applications, 2019. 128: p. 201- 213.

7. Zhang, C., et al., Detecting fake news for reducing misinformation risks using analyticsapproaches. European Journal of Operational Research, 2019.

8. Bondielli, A. and F. Marcelloni, A survey on fake news and rumour detection techniques. Information Sciences, 2019. 497: p. 38-55.

9. Ko, H., et al., Human-machine interaction: A case study on fake news detection using abacktracking based on a cognitive system. Cognitive Systems Research, 2019. 55: p. 77-81.

10. Zhang, X. and A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 2019.

11. Robbins, K.R., W. Zhang, and J.K. Bertrand, The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification. Journal of Mathematical Medicine and Biology, 2008

12. Alirezaei, M., S.T.A. Niaki, and S.A.A. Niaki, A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vectormachines. Expert Systems with Applications, 2019. 127: p. 47-57.

13. Zakeri, A. and A. Hokmabadi, Efficient feature selection method using real-valued grasshopper optimization algorithm. Expert Systems with Applications, 2019. 119: p. 61-72.

14. Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. News in an online world: Theneed for an "automatic crap detector". Proceedings of the Association for Information Science and Technology, 52(1):1–4.
15. Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4.
16. Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of NAACL-HLT, pages 7–17.
17. Balmas, M., 2014. When fake news becomes real: Combined exposure to multiple newssources and political attitudes of inefficacy, alienation, and cynicism. Communication Research 41, 430–454.
18. Pogue, D., 2017. How to stamp out fake news. Scientific American 316, 24–24.
19. Aldwairi, M. and A. Alwahedi, Detecting Fake News in Social Media Networks.Procedia Computer Science, 2018. 141: p. 215-222.
20. Mehdi H.A, Nasser G.A, Mohammad B, Text feature selection using ant colonyoptimization, Expert Systems with Applications, 2009

21. Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters,31(8), pp.651-666.

22. Quanquan Gu, Zhenhui Li, and J. Han, Generalized Fisher Score for Feature Selection.In: Proceedings of the International Conference on Uncertainty in Artificial Intelligence,2011

23. Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297. CiteSeerX

24. Reis, J.C., Correia, A., Murai, F., Veloso, A., Benevenuto, F. and Cambria, E., 2019. Supervised Learning for Fake News Detection.