WEST END UNIVERSITY COLLEGE

FAKE NEWS DETECTION USING UNSUPERVISED K-MEANS CLASSIFICATION ALGORTITHM

2022

WEST END UNIVERSITY COLLEGE

FAKE NEWS DETECTION USING UNSUPERVISED K-MEANS CLASSIFICATION ALGORTITHM

BY

Project submitted to the Faculty of Computer Studies, West End University College, in partial fulfillment of the requirement for the award of Bachelor of Science in Computer Science

JULY

2022

# DECLARATION

Candidate’s Declaration I hereby declare that this project is the result of my original study. No part of it has been produced for another degree in any university or elsewhere.

Candidate’s signature: ……………………… Date: ………………………………

Name:

Candidate’s signature: ……………………… Date: ………………………………

Name:

Supervisor’s signature:…………………………. Date: …………………………………..

Name:

## ABSTRACT

Language and written writings are the cornerstones of human communication. Email and text messages are also regarded as two of the primary sources of textual data, but social media is an essential source of information on the Internet. Text mining techniques are used to process and analyze text data. Text mining is the use of data mining to text files in order to identify patterns and extract pertinent information from massive amounts of text data. As major social media platforms like Facebook and Twitter acknowledged, there is a lot of bogus content, phony likes, views, and duplicate accounts. The majority of information shared on social media is dubious and occasionally inaccurate. To prevent a detrimental effect on society, they must be discovered as soon as feasible. The dimensions of the fake news databases are expanding quickly, so the dimensions must be decreased to improve false information detection while requiring less processing and complexity. Using the feature selection method is one of the greatest ways to reduce the quantity of the data. This method seeks to enhance classification performance by selecting a feature subset from the original set. One of the most essential text mining techniques is cluster analysis. Its objective is to automatically divide a set of objects into a finite number of homogeneous groups (clusters). In this thesis, we suggest using the K-means clustering algorithm to separate fake news from legitimate news coming from social media. The False News Detection project's objective is to see how machine learning and natural language processing can help with the fake news problem.

## DEDICATION

This project is dedicated to Mr. Alidu Abubakari , Lecturer, Faculty of Computer Studies, and West End University College.

Table of Contents

[DECLARATION 2](#_Toc110237467)

[ABSTRACT 3](#_Toc110237468)

[DEDICATION 4](#_Toc110237469)

[CHAPTER ONE 9](#_Toc110237470)

[Background of the Study 9](#_Toc110237471)

[Problem Statement 10](#_Toc110237472)

[Aim 10](#_Toc110237473)

[Objectives of the study 11](#_Toc110237474)

[SIGNIFICANCE OF THE STUDY 11](#_Toc110237475)

[SCOPE OF THE STUDY 11](#_Toc110237476)

[Fake News Detection 11](#_Toc110237477)

[Automated Fake News Detection 12](#_Toc110237478)

[An Approach Using Machine Learning 14](#_Toc110237479)

[K-Means Clustering Algorithm Benefits 16](#_Toc110237480)

[Limitation Of the Study 17](#_Toc110237481)

[CHAPTER TWO 18](#_Toc110237482)

[INTRODUCTION 18](#_Toc110237483)

[THEORY OF FAKE NEWS DETECTION 22](#_Toc110237484)

[CONTRIBUTORS OF FAKE NEWS 23](#_Toc110237485)

[LITERATURE REVIEW 24](#_Toc110237486)

[CHAPTER THREE 27](#_Toc110237487)

[METHODOLOGYs 27](#_Toc110237488)

[INTRODUCTION 27](#_Toc110237489)

[DATASET 28](#_Toc110237490)

[SIMULATION TOOL 31](#_Toc110237491)

[DATA ENTRY 38](#_Toc110237492)

[DATA PROCESSING 38](#_Toc110237493)

[In order to run machine learning or deep learning algorithms on text data, specific preprocessing is required. Text data can be transformed using a variety of techniques into a format that can be utilized for modeling. The news stories' headlines and content are both subjected to the data preparation processes outlined below. We also discuss the various word vector formats that we employed during our study. 38](#_Toc110237494)

[STOP WORD REMOVAL 39](#_Toc110237495)

[PUNCTUATION REMOVAL 40](#_Toc110237496)

[STEMMING 40](#_Toc110237497)

[TF-IDF vectorizer 42](#_Toc110237498)

[N-Grams 44](#_Toc110237499)

[Unigram 44](#_Toc110237500)

[Bigrams 44](#_Toc110237501)

[Important Factors to consider While Using K-means Algorithm 45](#_Toc110237502)

[Step 1: Initialization 46](#_Toc110237503)

[Step 2: Cluster Assignment 47](#_Toc110237504)

[Step 3: Moving Centroid 48](#_Toc110237505)

[Step 4: Optimization 49](#_Toc110237506)

[The aforementioned two processes are repeated until the centroids stop moving, that is, until they cease to vary their positions and become static. The k-means algorithm is said to have converged once this is completed. 49](#_Toc110237507)

[Step 5: Convergence 49](#_Toc110237508)

[Importing libraries 50](#_Toc110237509)

[Load data 51](#_Toc110237510)

[Data Preprocessing 51](#_Toc110237511)

[Merging of fake and True data frame 51](#_Toc110237512)

[Merging of fake and True data frame 52](#_Toc110237513)

[Word2Vec 52](#_Toc110237514)

[Sentence Vectors 52](#_Toc110237515)

[Clustering Algorithm 53](#_Toc110237516)

[Choice of K 53](#_Toc110237517)

[CHAPTER FOUR 57](#_Toc110237518)

[SIMULATION AND DISCUSSION 57](#_Toc110237519)

[CLUSTER VISULATION 57](#_Toc110237520)

[Accuracy 58](#_Toc110237521)

[DISCUSSION 58](#_Toc110237522)

[ACCURACY 58](#_Toc110237523)

[Our main measure for this experiment is K means which obtained 87% accuracy in classifying fake. Although other models perform better than K means, the clustering algorithm has an added advantage of relatively simple to implement and also the fact that it can scale to Scales to large data sets. 59](#_Toc110237524)

[CHAPTER FIVE 60](#_Toc110237525)

[SUMMARY, CONCLUSION AND RECOMMENDATION 60](#_Toc110237526)

[REFERENCES 61](#_Toc110237527)

[APPENDIX 66](#_Toc110237528)

[Figure 1 The machine learning process 21](#_Toc86052124)

[Figure 2 Example fo google colab 22](#_Toc86052125)

[Figure 3 Sample of training dataset 23](#_Toc86052126)

[Figure 4 Real and Fake news dataset 24](#_Toc86052127)

[Figure 5 Data Preprocessing 26](#_Toc86052128)

[Figure 6 Example of a stop word removal 27](#_Toc86052129)

[Figure 7 example of punctuation removal 27](#_Toc86052130)

[Figure 8 Example of steaming 28](#_Toc86052131)

[Figure 9 Example of bag of words 29](#_Toc86052132)

[Figure 10 Example of TF-IDF 30](#_Toc86052133)

[Figure 11 Example of n-grams 31](#_Toc86052134)

[Figure 12 Example of bigram and unigram 32](#_Toc86052135)

[Figure 13 Example of a naive bayes 36](#_Toc86052136)

[Figure 14 Example of gradient boost classifier 36](#_Toc86052137)

[Figure 15 Example of random forest classification 37](#_Toc86052138)

[Figure 16 Confusion Metrics 39](#_Toc86052139)

[Figure 17 Confusion Matrix for multinominal Naive Bayes 42](#_Toc86052140)

[Figure 18 Confusion Matrix on Gradient Boosting 42](#_Toc86052141)

[Figure 19 Confusion Matrix for Random Forest 43](#_Toc86052142)

[Figure 20 Accuracy result of all Models 44](#_Toc86052143)

# CHAPTER ONE

## Background of the Study

A growing number of people prefer to seek out and consume news from social media outlets rather than traditional news organizations as more and more of our lives are spent communicating online through social media platforms. These social media networks' changes, and their causes. Comparing social media to traditional news sources like newspapers and television, it is frequently timelier and less expensive to consume news there. It is also simpler to share, comment on, and discuss the news with friends or other users on social media. For instance, 62 percent of American adults reported receiving news via social media in 2016, compared to 49 percent in 2012. According to research, social media has surpassed television as the primary news medium. In 2003, Eric Schmidt, the former CEO of Google, stressed that "we create as much information currently every two days as we did from the birth of civilization up until 2003. Despite the benefits they offer, social media's news coverage is of poorer quality than that of traditional news sources. However, more fake news, or news pieces with purposefully incorrect material, is generated online for a variety of reasons, such as financial and political benefit, because it is inexpensive to provide news online and much faster and easier to disseminate through social media.

False news has the potential to seriously harm both people and society. A user finds it harder and harder to obtain information that is pertinent to their needs as there is a rise in the amount of material available. The ecosystem of news might become unbalanced due to fake news. It is clear that during the 2016 U.S. presidential election, the most popular fake news was even more widely disseminated on Facebook than the most popular mainstream news. Consumers are purposefully led to adopt prejudiced or erroneous opinions via fake news. Propagandists used it to spread political messages or exert political influence. As an illustration, some reports indicate that fake news was simply produced to incite people's mistrust and confusion.

Researchers have a lot to gain from using technologies like artificial intelligence (AI) and natural language processing (NLP) techniques to create systems that could recognize fake news automatically.

Text Mining (TM) algorithms are used to group, filter, and extract knowledge from the vast amount of digital texts. The Knowledge Discovery in Databases (KDD) technique is extended by text mining, which has a number of uses include classifying news articles based on their content, filtering email, and grouping documents or web pages. Language translation and issue analysis for customer support are further use cases.

## Problem Statement

Fake news can take many different forms, such as unintentional mistakes made by news aggregators, overtly fake stories, or tales created with the purpose to deceive and sway readers' opinions. Because it contradicts reality, it may generally have a detrimental impact on people, the government, and organizations.

Social media has gained popularity for news consumption over the past ten years because to its low cost, simple access, and incorrect information transmission. Social media, on the other hand, also makes it possible for "fake news," or news that contains purposefully misleading information, to spread widely. Social media fake news has lately emerged as a field of study that is receiving a lot of interest.

## Aim

To investigate the effectiveness of K mean clustering algorithm to detect fake news

## Objectives of the study

1. Create a method for cleaning the dataset as [art of the pre-processing.
2. Evaluating K means clustering algorithm as a means of classifying false news
3. Using several measures to assess the algorithm's performance

## SIGNIFICANCE OF THE STUDY

To recognize false news, we will rely on existing factual sources and news content aspects, therefore it will be knowledge- and style-based.

Platforms for social media have a lot of influence. Internet live statistics predict that there will be about 500 million tweets per day. These platforms can be found anywhere. They serve as the ideal forum for expressing thoughts, feelings, opinions, and intentions. This creates the ideal environment for news dissemination with the fewest restrictions and limitations.

Today's society is accustomed to getting news from online sources like social media. News is frequently viewed subjectively by readers. We frequently choose to consume things that trigger one of our many emotions. As a result, the news that is most widely reported may not be genuine or accurate. True news may sometimes be misrepresented while being transmitted.

## SCOPE OF THE STUDY

### Fake News Detection

Even defining fake news presents a hurdle. Fake news is a form of yellow journalism that consists of deliberate misinformation or hoaxes spread via traditional and broadcast news media or online social media. It can be described as a made-up story with the intent to deceive or news articles that are intentionally and verifiably false, and could mislead readers.

Finding the language (a collection of words or sentences) that is being used to fool the reader is the main task in the detection of false news. User-generated content can easily reach a large audience thanks to the social media sites' rising popularity. Social media has thus developed into a prime location for the spread of bogus news. For instance, during the 2016 U.S. Presidential election, the most talked-about false news tended to support Donald Trump over Hillary Clinton, causing significant societal harm and economic costs (Silverman, 2016). Due to the influence of fake news, some analysts have claimed that Donald Trump would not have won the presidency (Allcott & Gentzkow, 2017).

Identifying fake news frequently requires making use of all methods and processes available to verify the data. It might involve visiting websites for fact-checking. It can involve crowdsourcing verified news to contrast it with unverified news. However, the amount of data that is gathered every day on the internet is astounding. The speed at which information spreads online makes manual fact-checking feasible.

### Automated Fake News Detection

Scalability and automation are advantages of automated detection systems. Numerous tactics and strategies are used in fake news detection research. It's also important to note that, depending on one's perspective, these strategies commonly cross over. There are just two strategies, in my opinion, that are worth talking about. These two strategies place more emphasis on the procedures employed than the subject matter under investigation. Both of their methods may make use of natural language processing (NLP).

Computers can understand and respond to natural/human language thanks to natural language processing. As a result, there are two factors to take into account:

* Natural Language Processing (NLP)

The term "computers understanding the structure and meaning of human language" (such as English, Spanish, and Japanese) is used by Gartner to describe Natural Language Understanding (NLU), which enables humans to interact with computers using everyday expressions. In other words, NLU is a type of artificial intelligence that uses software to analyze text and other unstructured data. A text can be ingested by NLU, converted to computer language, and then output in a human-readable form.

* Natural Language Processing

NLG is a software method that automatically transforms data into content in plain English. It belongs to the artificial intelligence sector (AI). The system may actually build a story that is identical to that of a human analyst by creating the words and paragraphs for you. One of the business sectors that is gaining ground the quickest is NLG. NLG has many uses, but it works best when it is automated to handle time-consuming data processing and reporting duties.

The two methods for identifying fake news are:

1. Machine Learning
2. Deep Learning

### An Approach Using Machine Learning

Machine learning is the process of enabling computers to learn without being explicitly programmed. Numerous classification algorithms, including decision trees, logistic regression, support vector machines, and naive Bayes classifiers, are available thanks to data science. The random forest classifier, however, is near the top of the classifier hierarchy. A machine learning technique using machine learning algorithms to find false information. Some instances of these algorithms are as follows:

* Multinomial naive bayes

Naive Bayes is based on the Bayes theorem, which claims that features in a dataset are independent of one another. The likelihood of one attribute occurring does not affect the likelihood of the other. For tiny sample sets, Nave Bayes can outperform the most potent alternatives. It is used in many different disciplines because of its relative durability, simplicity, efficiency, and precision.

* Random Forest Classification

A supervised learning algorithm is random forest. An ensemble of decision trees, typically trained using the "bagging" approach, make up the "forest" that it constructs. The bagging method's general premise is that combining learning models improves the end outcome. Simply put, random forest creates many decision trees and integrates them to get a prediction that is more accurate and reliable. Random forest has the key benefit of being applicable to both classification and regression problems, which make up the majority of modern machine learning systems.

* Gradient Boost Classification

Final predictions are produced by the Gradient Boosting Machine (GBM), which combines forecasts from numerous decision trees. Remember that all of the weak learners in a gradient boosting machine are decision trees.

Datasets are used to refine the algorithms. Training data and test data are two categories into which these datasets can be separated. Numerous studies combine data mining and machine learning methods. On social media platforms, notably Twitter, this is a standard practice. To identify fake news, a model can, for instance, use Nave Bayes, Support Vector Machines (SVM), and Natural Language Processing (NLP).

Depending on the type of data, the two classifiers can be applied to a dataset and their performance compared. On the other hand, these classifiers can be used in an ensemble technique to boost each other's performance in classification tasks, hence increasing model accuracy.

To extract features from data, NLP may be employed. Additionally, since contextualizing text input is a challenge for standard machine learning algorithms, it might be helpful in those situations. NLP can be used to analyze the sentiment of data since sentiment analysis is one of its subfields.

With the help of this thesis, it will be possible to automate the detection of fake news by determining which features work best for K means classifier. We'll evaluate how well various extracted attributes work for detecting fake news. Features from the articles must be extracted for the classifiers to be trained on when classifying text using machine learning methods. Word counts, ngram counts, term frequency-inverse document frequency, sentiment analysis, lemmatization, and named entity recognition are just a few of the features that are extracted from the dataset.

## K-Means Clustering Algorithm Benefits

The advantages are listed below:

* Fast and reliable
* Simple to comprehend
* Comparably effective
* The best outcomes are obtained when data sets are distinct.
* Create more compact clusters.
* The cluster varies as centroids are recomputed.
* Flexible
* Better computational cost, Simple to interpret
* Enhances Accuracy
* uses spherical clusters better

***Disadvantages of K- Means Clustering Algorithm***

***Below are the disadvantages:***

* Needs to be specified in advance how many cluster centers there will be.
* When two sets of data are heavily overlapping, it is impossible to identify that there are two clusters.
* The outcomes are different as a result of the various data representations.
* The factors may not be equally weighted by Euclidean distance.
* It provides the squared error function's local maxima.
* Sometimes picking the centroids at random won't produce good results.
* The computer could crash if it encounters particularly huge data sets.
* predicting problems

## Limitation Of the Study

Constraints on the problem include the irregularity of the data, which makes it possible for any prediction model to be inaccurate.

# CHAPTER TWO

## INTRODUCTION

The world is evolving quickly. Contrary to traditional mass media like newspapers, periodicals, radio, and television, the growth of social networks has led to an epidemic of fake news (de Oliveira, Pisa, Lopez, de Medeiros, & Mattos, 2021). In this digital age, there are undoubtedly many diverse problems. The fake news is one of them. The ease of access, cheap cost, and quick dissemination of information that social media offers encourage consumers to seek out and consume news there. The benefit is that it makes it easier for "fake news," or low-quality news with blatantly erroneous content, to proliferate widely. The dissemination of false information could have a severe effect on people and society.

Online sources for news include search engines, social networking sites, and the homepages of news organizations. On the other hand, manually determining the veracity of news is a challenging process that typically calls for topic specialists who rigorously examine assertions, supporting evidence, context, and reporting from sources that you can trust Annotated news data is frequently sourced from professional journalists, fact-checking websites, industry detectors, and crowdsourcing personnel.

The challenge of identifying false information sources through content-based analysis is believed to be solvable, at least in the context of spam detection. To assess whether content (such as tweets or emails) is spam or not, spam detection uses statistical machine learning algorithms. These techniques include text pre-processing, feature extraction (i.e., a bag of words), and feature selection based on which features produce the best results on a test dataset. Then, these features can be categorized using classifiers like K-nearest neighbors, Nave Bayes, Support Vector Machines, and TF-IDF. All of these classifiers are supervised since they are supervised machine learning classifiers.

where is the message to be categorized and is a parameter vector, and and are spam and legitimate messages, respectively.

The issue of identifying fake news is related to and almost analogous to the task of identifying spam in that both aim to distinguish examples of legitimate text from examples of illegitimate, poorly intended messages. How we might use similar methods to identify fake news becomes the next difficult task. Instead of filtering spam, it would be beneficial to be able to identify fake news stories and flag them, letting readers know that what they are reading is probably false.

This problem has been approached by numerous researchers in a variety of methods to see which approach is most effective and produces the best outcomes. A few papers have examined feature extraction and model building as false news detection techniques from a data mining perspective. However, the issue is not that straightforward. A methodology of feature extraction (both news content features and social context features combined with metric evaluation using precision, recall, and f1 scores) has proven to bear educated results.

It can be challenging to find and categorize news that has been published online in an unstructured format (such as news, articles, videos, and audios), as this strictly requires human skill. However, anomalies that distinguish text articles that are deceptive in character from those that are based on facts can be found using computer techniques like natural language processing (NLP).

We need to integrate auxiliary information, such as user social engagements on social media, to aid make a conclusion because fake news is purposefully designed to deceive readers into believing false information, making it challenging and nontrivial to identify based on news content. Second, making use of this supplementary data is difficult in and of itself due to the large, sparse, chaotic, and noisy data that users' social interactions with fake news produce (Shu, Amy, Suhang , Tang, & Huan)

The cousins of fake news are not new. The "Great Moon Hoax" of 1835, in which the New York Sun published a number of articles regarding the discovery of a lie on the moon, serves as one historical illustration. A contemporary instance is the "Flemish Secession Hoax" of 2006, in which a Belgian public television station falsely claimed that the Flemish parliament had proclaimed independence from Belgium. Many viewers took this news to be accurate. Supermarket tabloids like the Weekly World News and the National Enquirer have long been known to publish a mixture of articles that are both partially factual and downright fake. (Allcott & Gentzkow, 2017)

The interface has the potential to separate useful information from online services, especially as news becomes important for decision making. Clickbait seduce people and pique their interest with eye-catching headlines or design to click links to enhance advertisement earnings (Aldwairi & Alwahedi, 2018). The language that computers can understand is called machine language. The only thing that a computer can comprehend is this. Machine language is the end result of all programs and programming languages (Schmit, 2015).

Humans are prone to Truth-Bias, Naive Realism, and Confirmation Bias, which is the main cause of the persistence of false information. People who are naturally predisposed in favor of the truth are said to have "the presumption of truth" in their favor. "The propensity to evaluate someone based on their social encounters" This presumption is founded on the idea that the interpersonal message is genuine and is only likely to change if the situation changes and raises suspicion (V, 2017).

Humans are, in general, terrible lie detectors, and the risk that they are being tricked increases when they are unaware that there is a problem. The majority of social media users are not aware that there are posts. Tweets, articles, or other written materials that use the phrase "single objective of influencing others' beliefs in order to influence others' views in order to influence others' beliefs in order to influence others' beliefs to modify the way they make decisions" Data manipulation is unlawful. It's a poorly understood topic that's frequently ignored, especially when erroneous information is conveyed. When a buddy recommends something, users often let their guard down and can take all of the false information posted on social media as gospel. This is particularly harmful given that young users frequently rely on social media to keep up with politics, significant events, and breaking news. (V, 2017).

Additionally, Naive Realism, which is the tendency for people to think that their own opinions are the only ones that are valid, is used to describe those who hold different opinions as being "uniformed, unreasonable, or biased (Shu, Amy, Suhang , Tang, & Huan)."

This raises the issue of confirmation bias, which is the idea that people prefer learning things that only confirm their current beliefs. Customers do not want to find any proof that contradicts their beliefs; they just want to hear what they believe. For instance, someone who strongly favors unfettered gun control might want to use all evidence they can find to advance their arguments and justifications.

The normal individual, who is uninformed of erroneous information to begin with, will not be able to resist these inadvertent tendencies. Only those that aim for particular academic standards may be able to avoid or restrict any biasness.

Fake news has a negative impact on people, but it also has long-term bad effects on society. Fake news is capable of upsetting the "balance of the news ecosystem" because of all the misleading information that is out there (Shu, Amy, Suhang , Tang, & Huan).

## THEORY OF FAKE NEWS DETECTION

Knowledge-based: Checking the veracity of a new article's primary claims to determine the news's veracity is the simplest way to identify fake news. The purpose of this is to leverage other sources to verify assertions made in news articles. Assigning a truth value to a claim in a certain situation is the objective. Fact-checking methods currently in use can be divided into three categories: expert-oriented, crowdsourcing-oriented, and computational-oriented.

Expectation-oriented fact-checking uses human specialists to examine pertinent data and documents and develop judgments about the truthfulness of claims. However, this requires a lot of time and mental effort.

* Fact-checking focused on crowdsourcing uses the "wisdom of crowd" to let regular people annotate news information. These observations were then combined to create a comprehensive evaluation of the news's authenticity.
* • Computationally oriented fact-checking seeks to offer a scalable, automated system to classify true and fraudulent statements. Prior computation-oriented fact-checking techniques aim to address two main problems; recognizing verifiable claims and determining whether factual claims are true.
* Style-based: These techniques aim to identify fake news by identifying the manipulators in the news content's writing style. Style-based approaches can be divided into two general categories: deception-oriented and objectivity-oriented.
* Deceptive assertions or claims from news content are captured by deception-oriented stylometric approaches. The Undeutsch Hypothesis serves as the inspiration for deception detection, and numerous forensic tools, such as criteria-based content analysis and scientific-based content analysis, have been developed.

Objectivity-oriented techniques identify stylistic cues, such as hyperpartisanship and yellow journalism that may point to a loss of objectivity in news reporting and the possibility for consumer deception. Extreme political advocacy is characterized by hyperpartisan styles, which frequently coincide with a strong drive to spread false information. Hyperpartisan publications can be found using linguistic-based characteristics.

## CONTRIBUTORS OF FAKE NEWS

While many social media users are definitely actual people, there is no way of knowing for certain whether those who are nasty and trying to spread lies are real people. There are three primary categories of sources for false news: Robots, trolls, and cyborgs may use the internet to spread lies and half-truths. It is crucial to comprehend them in order to comprehend how the internet spreads false information.

Social media is increasingly being used to promote false information and polarize Americans over divisive issues like race and immigration, as the 2016 election showed. Many of them are harmless, using Twitter to send strange poems or pictures of pets. On the other side, there are some who are up to naught good and are only pretending to be individuals. Due to the inexpensive cost of opening a social media account, the establishment of destructive accounts is not discouraged. Even the best researchers often struggle to identify bots.

## LITERATURE REVIEW

There is a solid framework for future systems to build on, but the methods for verifying news and facts and removing inaccurate content are now limited. These methods, which are typically implemented in response to user complaints, either make an independent attempt to verify content or entirely filter off suspect sources by using past knowledge and prejudices. For instance, businesses like PolitiFact and Snopes hire experts to manually evaluate claims, and institutions typically offer instructions for their students on how to identify reliable information sources. In an effort to stop the spread of an online fake news article, even Facebook turned to third-party fact-checking organizations like Snopes and PolitiFact to flag disputed tales (Leong, 2017).

The reader must conduct their own due diligence because it is not always possible to wait for journalists to conduct in-depth research considering how quickly spectacular news spreads.

Several different classifiers have been tried by researchers to identify fake news. Convolutional neural networks (CNN) and long short-term memory units (LSTM) are two examples (Kim, 2014).

Despite being created for machine vision applications, CNNs often have a decent level of effectiveness.

But occasionally, LSTMs outperform CNNs. LSTMs have the capacity to "forget" some information and concentrate on, or "remember," more important information. They therefore function well with substantial amounts of text data (Y. Long, 2017).

A classifier developed by Rubin et al. differentiated between "genuine news" and "satire news" with 90% accuracy (V. Rubin, 2016). They concentrated on "satire news" since while it is misleading, it does not have the same malicious intent as fake news. Satire, in contrast to Fake News, which aims to mislead, is supposed to be blatantly fake. Because satire is easier to identify than fake news, Rubin et al prefer to concentrate on it. There were only 290 training articles and 90 test articles in the dataset utilized by Rubin et al. They employ a Support Vector Machine as a classifier, which works well for binary classification but not for multiclass classification. (V. Rubin, 2016).

A Google research team made an effort in 2014 to create a system that could determine the veracity of information culled from the internet (Dong, 2014). They designed their research to address the challenge of determining the likelihood that a scraped subject-predicate-object triple of information is in fact true. The data fusion problem, which is the difficulty of determining the true values of data given a pool of conflicting information (e.g., identifying the birthplace of President Obama from a set of potential locations extracted from various articles, blogs, and editorials), builds on this problem, which they called the knowledge fusion problem.

The fact that prior research typically included all detection features, which results in significant computational complexity, is one of their key problems. Due to the detection method taking into account redundant irrelevant features, this also leads to low classification precision. High-dimensional datasets reduce the classifier's functionality in two ways: first, they need more processing, and second, because the models they enable have less generalizability, they are more prone to overfitting. Therefore, lowering the size of the datasets can simplify computations and enhance the effectiveness of the classification algorithms.

Although K-means is a well-known technique for data clustering in text mining, feature selection rarely uses it. Words that can accurately describe a class's semantics are typically good features for text data. We collect numerous cluster centroids for each class using the k-means approach, and we then select the high frequency words in the centroids as the text features for categorization. The words that were retrieved using k-means possess good quality for semantic expression in addition to being able to accurately reflect each class clustering.

# CHAPTER THREE

## METHODOLOGYs

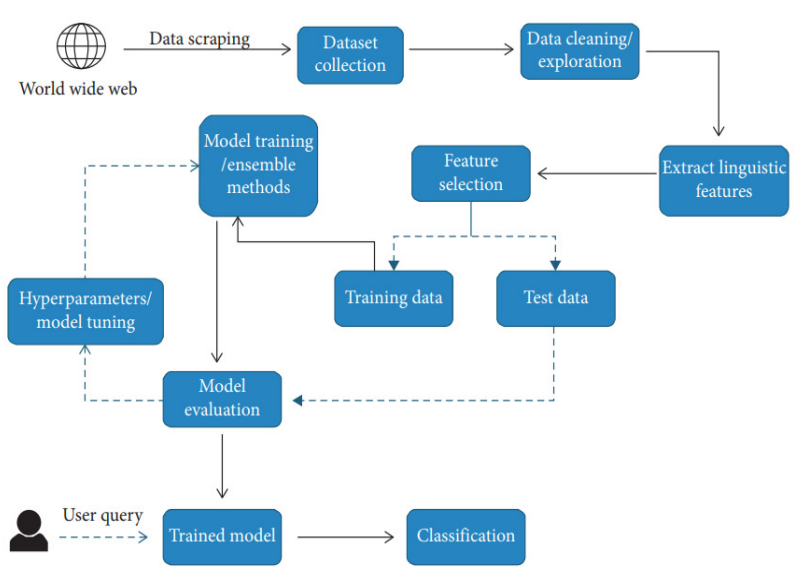


Figure 1 The machine learning process

## INTRODUCTION

The methodology used to accomplish the study's intended objectives is described in detail in this chapter. Figure 1 shows how machine learning has given computers a wide range of new abilities, from movie prediction to cancer analysis. The system that automates this process is called a "Model," and it is set up using a "Training" technique. We need to collect data in order to ask the questions we need to in order to create an accurate model.

Finding patterns that are associated with a piece of news that may be fraudulent is the aim of this research. It goes without saying that human assistance is necessary at some point in any categorization process. Even while perfect accuracy would not be feasible, identifying the characteristics of fake news would be a positive step. The idea is to build a model that would correlate a piece of news with the likelihood that it is fake news by first gathering a significant amount of data that has already been identified and validated as fake news.

It is necessary to transform the raw text data into something more helpful in order to categorize news stories. The term for this is feature extraction. Word counts, n-gram counts, punctuation usage, sentiment analysis, and many more techniques are all examples of feature extraction. The article from which the features were taken can then be classified using the extracted features. Depending on the underlying data patterns, various features may produce varying results. It is possible to identify patterns in the data by experimenting with various classifiers and various characteristics. It is possible to boost the likelihood of automated fake news identification by identifying the most useful features for categorizing fake news.

## DATASET

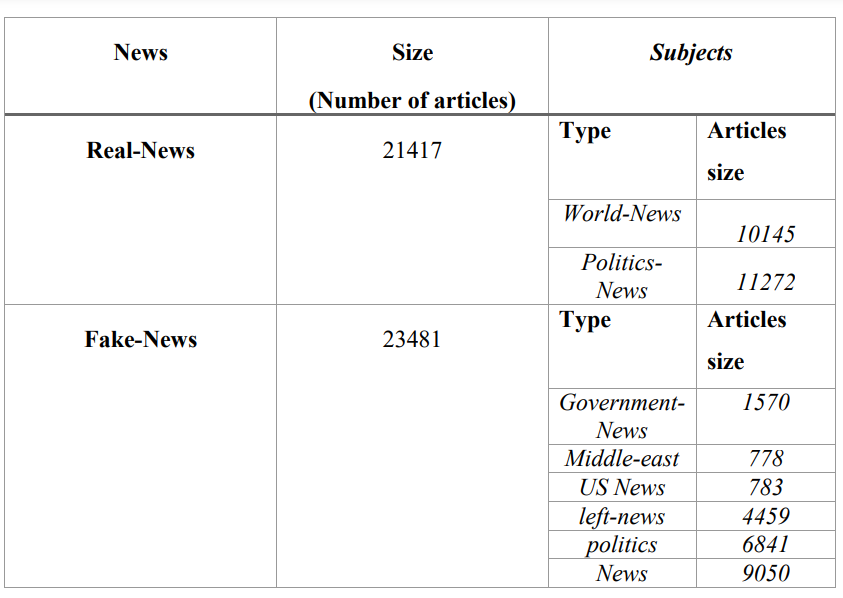
First news must be gathered and tagged in order to identify trends in fake news. There needs to be roughly equal representation of both real news and fake news. This prevents the frequency of fake news in the dataset from being a criterion for classification. Valid results must be produced with reliable data. Good data in this context is data that is generalizable and representational of the real world.

The ISOT Fake News Dataset, the largest collection of full-length fake news articles, was utilized to train the classifiers (H. Ahmed, 2017).

The dataset includes both actual and fraudulent news stories. The accurate articles were retrieved by crawling Reuters.com (a news website), which is how this dataset was compiled from real-world sources. The phony news pieces were gathered from a variety of sources. The false news reports were gathered from shady websites that Wikipedia and the American fact-checking group Politifact had identified as being unreliable. The dataset includes a variety of articles on various subjects, although the majority of them are about politics and global news.

There are two CSV files in the dataset. There are more than 12,600 items from reuter.com in the first file, "True.csv." The second file, "Fake.csv," has more than 12,600 articles from various sources used by fake news outlets. Each article includes the following details: the article's title, text, format, and publication date. We mostly collected articles from 2016 to 2017 in order to match the false news data gathered for kaggle.com. The data was cleaned and processed, but the text still had the punctuation errors and other errors seen in the bogus news.

The categories and amount of articles for each category are broken out in the following table.



80 percent of the ISOT data will initially be used to train each model. The trained classifiers' accuracy will be evaluated using the remaining 20% of the ISOT data. As previously noted, testing will also include FakeNewsNet and the Original Data. To ensure that we are identifying Fake News and not another pattern of the ISOT dataset, such as a style of a certain news agency, we are employing these extra tests.

Each article in the ISOT dataset marked as Real was obtained from Reuters; all of the stories there began with the term "Reuters." Both people and machines would be able to recognize this pattern with ease. The opening "Reuters" sentence was omitted from each item to avoid this problem.

## SIMULATION TOOL

Free to use, Google Colab is a cloud-based Jupyter notebook environment. Most significantly, it doesn't require any setup, and, like Google Docs pages, the notes you create can be edited simultaneously by team members. Colab supports a wide range of popular machine learning libraries, which you can rapidly load into your notebook.

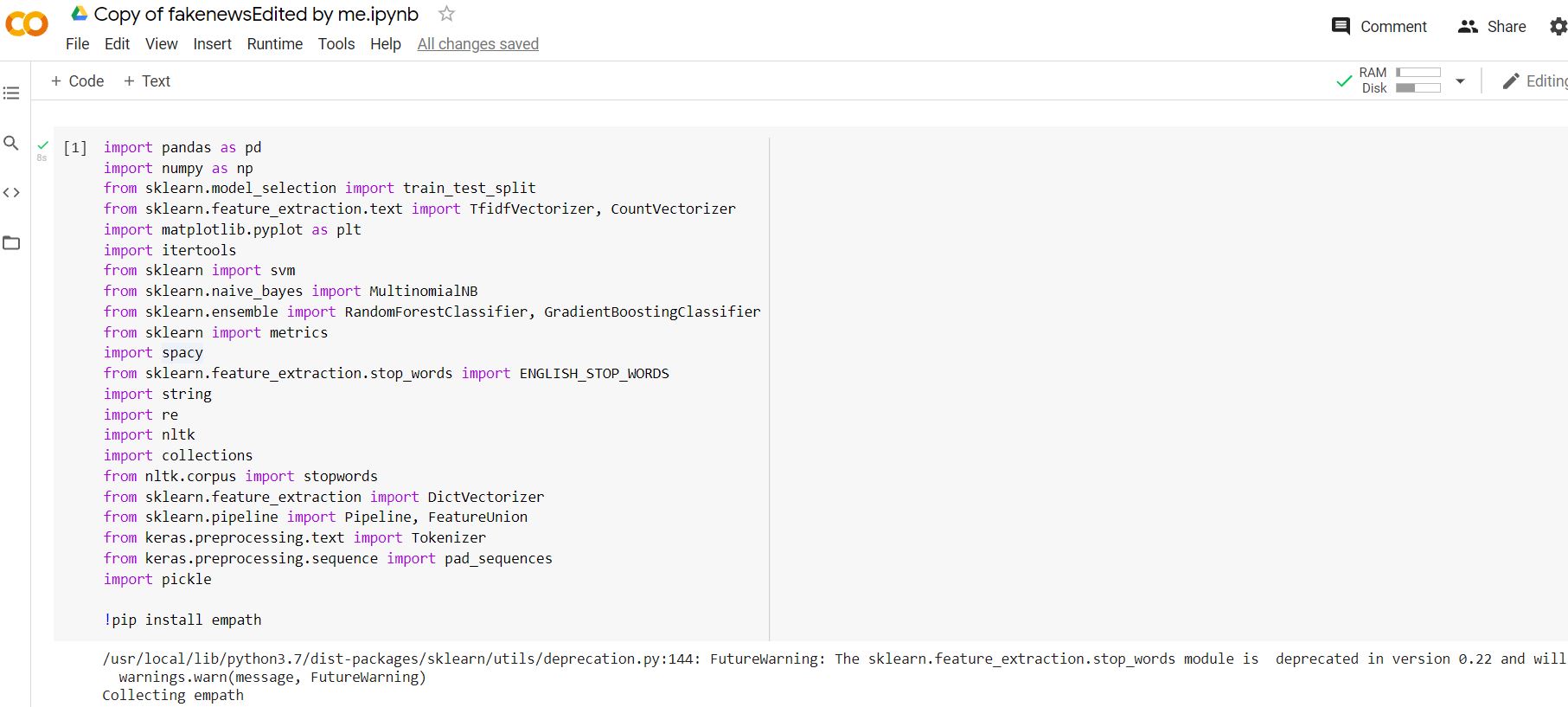


Figure 2 Example of google colab (colab, n.d.)

Thus the below figure 4.1 shows the sample input data of ISOT Fake news dataset

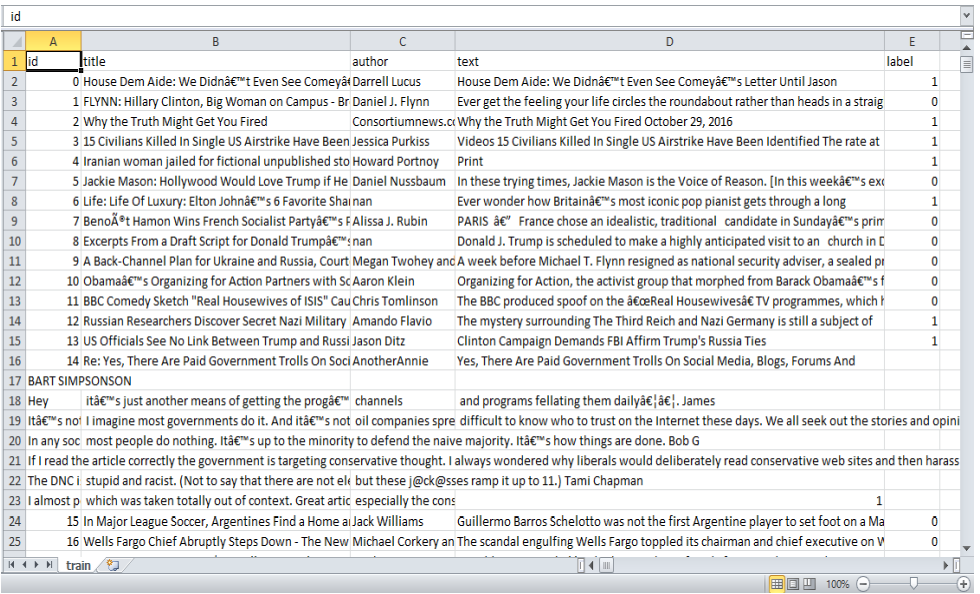


Figure 3 Sample of training dataset

**Data Exploration**

Figure 4 shows the distribution of both fake and real news the ISOT dataset. As can be observed the data is imbalanced.

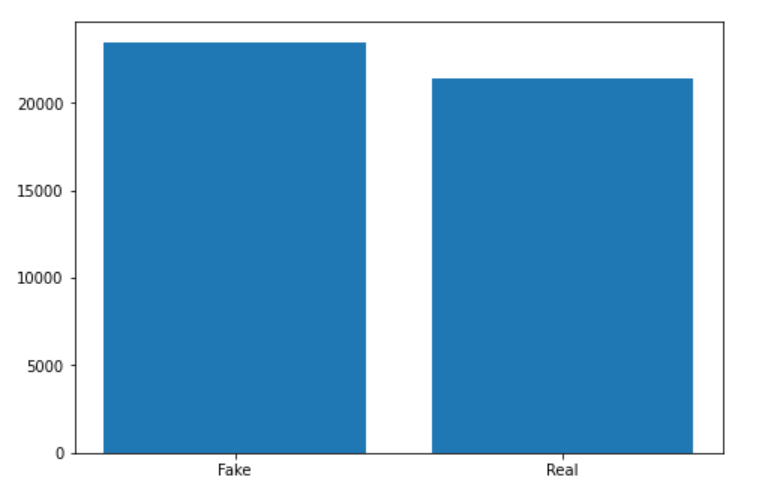


Figure 4 Real and Fake news dataset (colab, n.d.)

**Data Exploration (Fake News)**

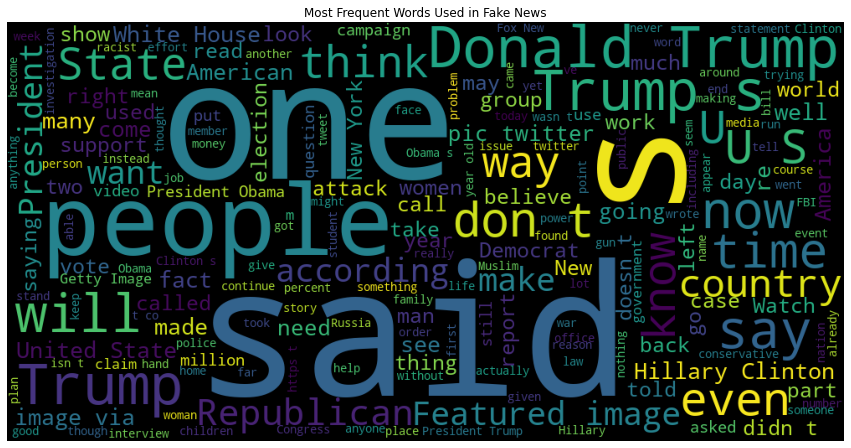
Word clouds were created for both true and fraudulent news items. It is evident that in fake articles (Figure 5), the frequency is evenly dispersed over many of the words with the exception of the phrase "Donald Trump," however this is not the case in real publications (Figure 6) Words like "Government," "Donald Trump," and "WASHINGTON Reuters" stand out from the others, as do "United States" and "White House." 

Figure 5 Most frequent words used in fake news

**Data Exploration (Real News)**

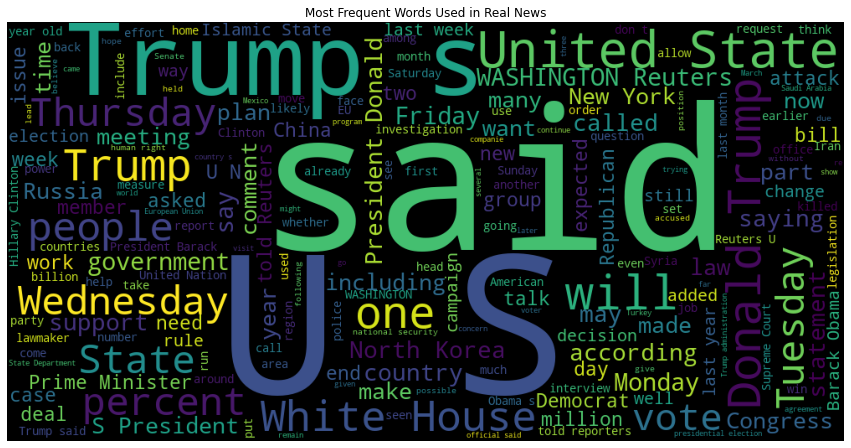


Figure 5 Most frequent words used in Real news

Data Exploration (Links & Tags)

100%|██████████| 44898/44898 [00:00<00:00, 71815.40it/s]

51%|█████ | 22769/44898 [00:00<00:00, 227670.13it/s]

Total URLS: 4672

100%|██████████| 44898/44898 [00:00<00:00, 221986.30it/s]

Total Tags: 246

Checking Null or Missing Values

|  |  |
| --- | --- |
| title | 0 |
| text | 0 |
| subject | 0 |
| date | 0 |
| Label | 0 |
| Dtype | Int64 |

**Histograms**

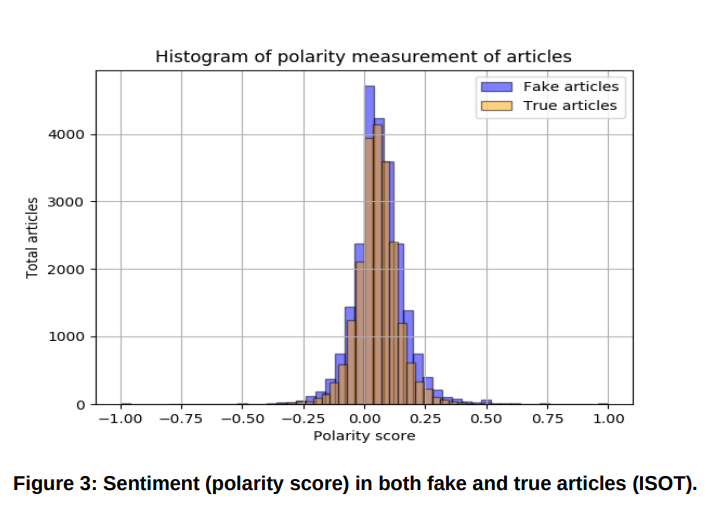
The linguistic characteristics (text length, word count), as well as the sentiment dimension, of the article text in the ISOT dataset were highlighted using histograms.

Figure 6 Popularity Score in both fake and true articles (ISOT) (Iatropoulou, 2020).

Since both real and false articles are similarly distributed around 0, the sentiment histogram (Figure 6) demonstrates that the corpus of the particular dataset is made up of neutral sentiment texts (Iatropoulou, 2020).

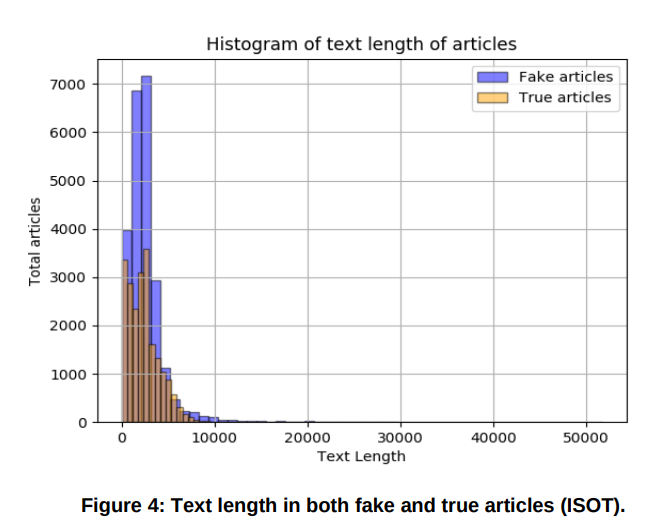


Figure 7 Text length in both fake and true articles (ISOT) (Iatropoulou, 2020).

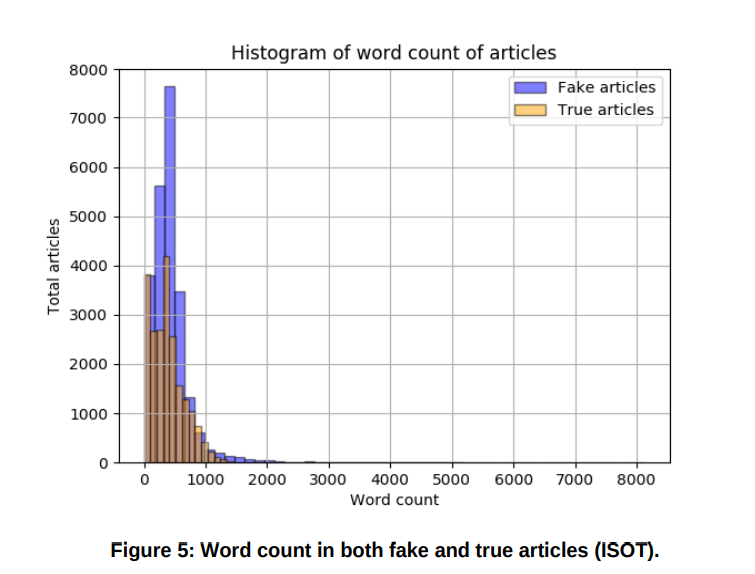


Figure 8 Word count in both fake and true articles (ISOT) (Iatropoulou, 2020).

The entire text length of both the real and false articles is between 10 and 1000 words (Figure 7). True articles can contain up to 1000 words (Figure 8), whereas false writings often contain 500–600 words, with some publications using as many as 2000 words.

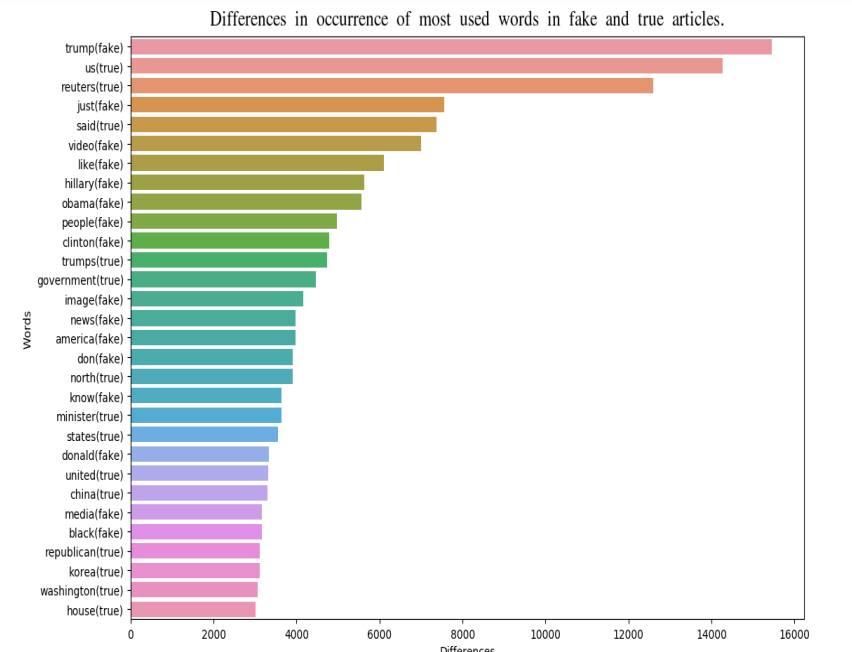


Figure 9 Differences in occurrence of most used words in fake and true articles (Iatropoulou, 2020).

By summing up the frequency of the frequent phrases in both types of articles and estimating the disparities, it seems instructive to comprehend which common terms could make a difference in recognizing an article as truthful or untrue. The word "trump," for example, is prevalent in both categories, as seen in Figures 6 and 7, but its occurrence in fake articles is more than 15000 times more frequent than in real publications. (Iatropoulou, 2020).

## DATA ENTRY

The datasets we used for this study are open source and freely downloadable online. The information comprises both fake and actual news articles from a variety of sources. Authentic news articles produced provide an accurate account of actual occurrences, in contrast to fake news websites that make claims that are not backed by the facts. With the use of fact-checking websites like politifact.com and snopes.com, many of the political assertions made in those publications may be manually confirmed.

Figure 10 Data Preprocessing

### DATA PROCESSING

## In order to run machine learning or deep learning algorithms on text data, specific preprocessing is required. Text data can be transformed using a variety of techniques into a format that can be utilized for modeling. The news stories' headlines and content are both subjected to the data preparation processes outlined below. We also discuss the various word vector formats that we employed during our study.

## STOP WORD REMOVAL

Stop words are first eliminated from the text data provided. Stop words, which are the most common terms in a language but offer no context, may be processed and filtered out of the text because they are more prevalent and contain less useful information. Instead of conjunctions like "and," "or," and "but," prepositions like "of," "of," "in," "from," and "to," and the articles "a," "an," and "the," stop words function more as a connecting element of the sentences. A important first step in natural language processing is the elimination of stop words because stop words that are less relevant may cause processing time to be wasted.

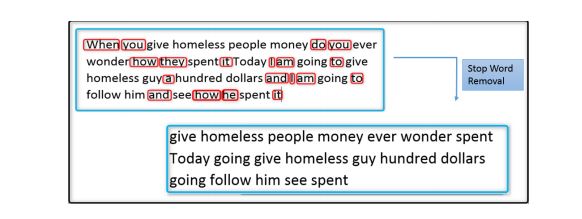


Figure 11 Example of a stop word removal (Lakshmikumar, 2019)

### PUNCTUATION REMOVAL

In everyday speech, punctuation provides the phrase with grammatical context. Commas and other punctuation marks may not significantly aid in understanding what the phrase means.

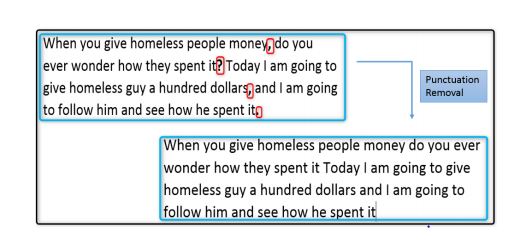


Figure 12 example of punctuation removal (Lakshmikumar, 2019)

## STEMMING

Stemming is the process of stripping a word of its prefixes and suffixes until just the stem is left. By employing stemming, we can condense inflectional and occasionally derivationally connected word variations to a basic form. The stemming procedure is shown in action in Figure 13.

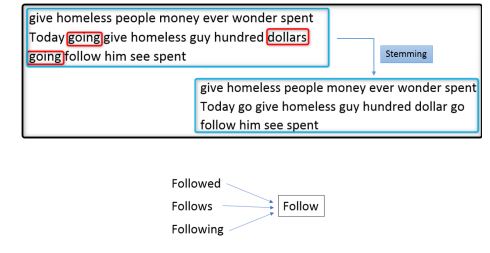


Figure 13 Example of steaming (Lakshmikumar, 2019)

**Word Vector Representation**

Preparing the text for modeling from the news article's body and title is challenging. To perform text analytics, we must convert unstructured text into numerical traits. We experimented with two distinct methods to transform the raw text and extract features: Bag of Words by TF-IDF.

**Bag of Word**

The Bag of Words (BoW) approach analyzes each news article as a document and determines the frequency count of each word in that document. From there, fixed-length vector features, which are numerical representations of the data, are produced. To create a word count vector from raw text for feature extraction, Bag of Words employs the CountVectorizer function. The text is separated from the material by CountVectorizer, which also develops a vocabulary and turns the text into a vector. This encoded vector will contain a count for each word's occurrences that resembles a frequency count as a key value pair. In terms of information loss, this strategy has drawbacks. The relative location of the words is not taken into account, and context information is lost. This loss could be expensive when compared to the gain in computational simplicity with the ease of use.

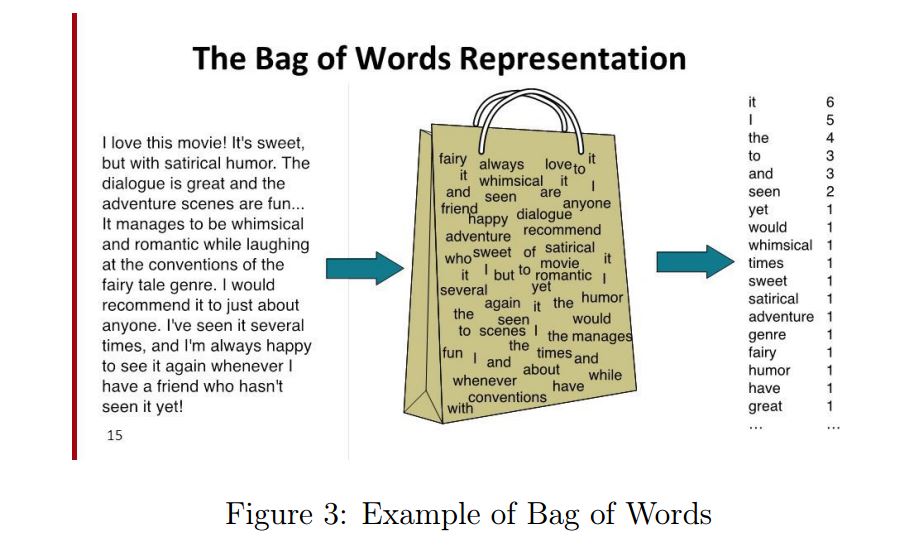


Figure 14 Example of bag of words

## TF-IDF vectorizer

We used the "Term Frequency-Inverse Document Frequency" (TF-IDF) approach for feature extraction. Term Frequency and Inverse Document Frequency are the two parts of the TF-IDF. The local significance of a word is determined by how frequently it appears in a document. Using Inverse Document Frequency, the signature words that do not appear more frequently across documents are found.

A signature word with a high TF-IDF is one that is crucial for the given document, occurs frequently there, but is uncommon elsewhere in texts.

1. Term frequency, or TF, refers to how frequently a word appears in a document. The formula for TF is: TF(w) = (Word 'w' Appearances in Document) / (Total number of words in the document).
2. IDF (Inverse Document Frequency): This gauges a word's significance inside the text. Words like and, of, the, and a, for instance, are frequently used but have less significance. As a result, less common terms are given greater weight while the most frequently repeated terms are given less weight. IDF(w) = loge (Total number of documents / Number of documents containing the term "w").

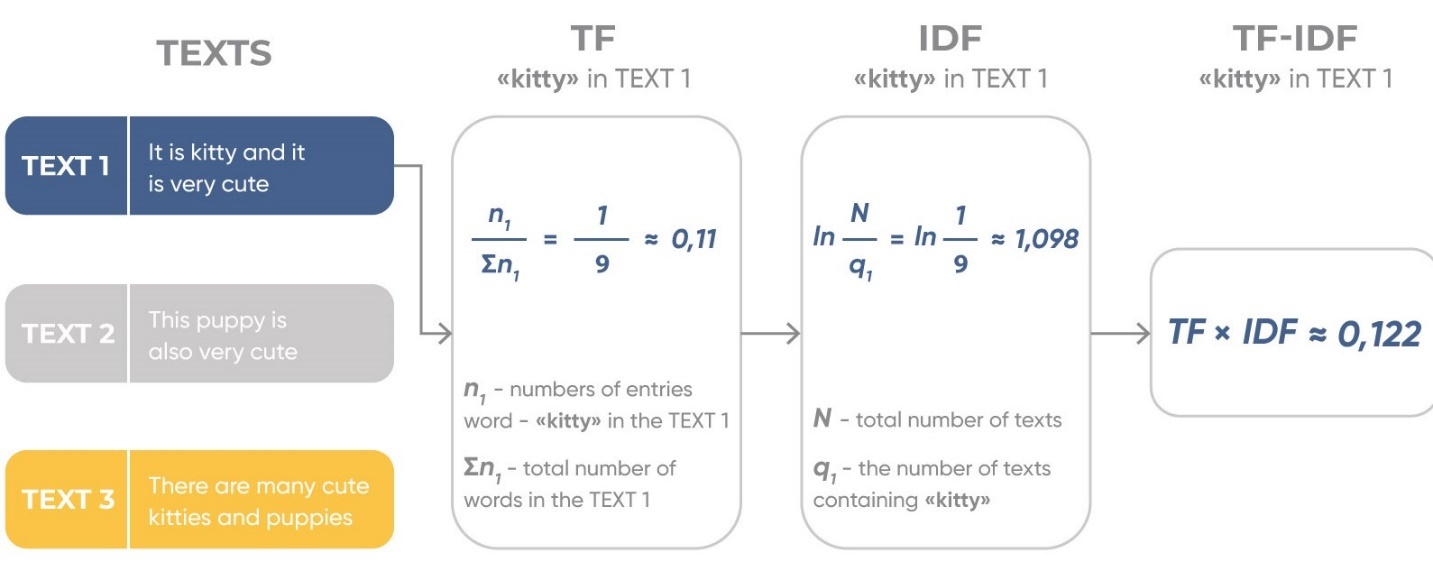


Figure 15 Example of TF-IDF (K.P.Singh, 2018)

The TF-IDF weight is given to each word by calculating TF\*IDF values.

number of occurrences of i in j

number of documents containing i

total number of documents

We determine the bigrams' values and express the TF-IDF vector of those bigrams in order to generate the news vector.

## N-Grams

An n-gram is a continuous string of n words taken from a particular text. A "bigram" is a size 2, a "trigram" is a size 3, and so on. A "unigram" is a size 1 n-gram. A larger n allows a model to hold more context.

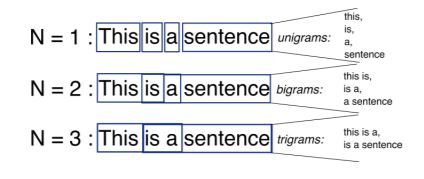


Figure 16 Example of n-grams (K.P.Singh, 2018)

## Unigram

A one-word sequence is known as a 1-gram (or unigram). I, "love," reading, blogs, about, data, science, on, analytics, vidhya — those would be the only unigrams.

## Bigrams

A bigram is any pair of consecutive elements in a string of tokens, which are typically letters, syllables, or words; they are also known as n-grams for n=2. The frequency distribution of bigrams in a string is frequently used in numerous applications, such as computational linguistics, encryption, and speech recognition, for fundamental statistical analysis of text. Word pairs with gaps between them are known as gappy bigrams, commonly referred to as skipping bigrams. Head word bigrams are gappy bigrams with an explicit dependent relationship. Bigrams are used to calculate a token's conditional probability given the previous token.

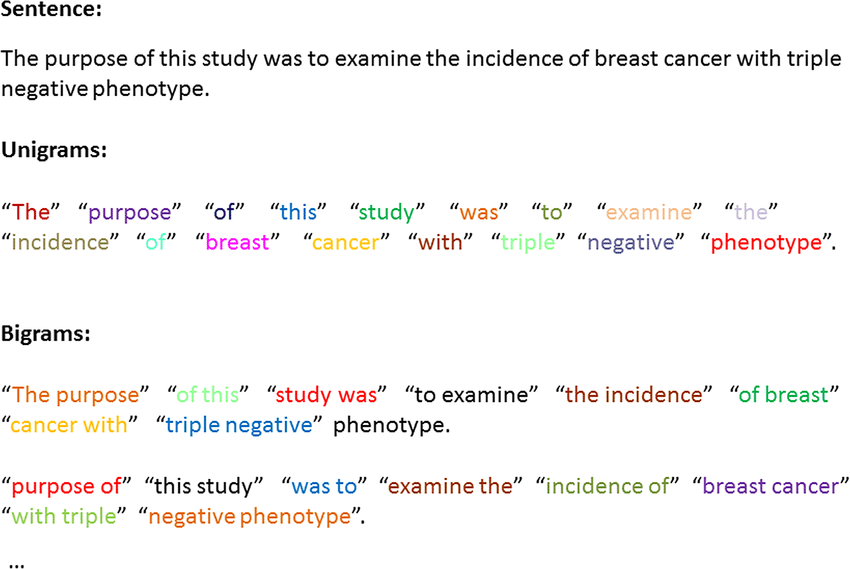


Figure 17 Example of bigram and unigram (Rastegar-Mojarad, 2015)

## Important Factors to consider While Using K-means Algorithm

The effectiveness of the final clusters created when employing k-means clustering can be affected by a number of things. Therefore, when employing the K-means clustering algorithm to solve business challenges, we must bear the following things in mind.

1. Number of clusters (K): You must specify how many clusters you wish to group your data points into.

2. First Values/Seeds: The cluster formation can be influenced by the initial cluster centers that are chosen. K-means is a non-deterministic algorithm. This means that clustering results, even when applied to the same data set, can vary from run to run.

3. Outliers: The existence of outliers has a significant impact on cluster formation. Outliers influence the best cluster formation by pulling the cluster toward itself.

4. Distance Measures: Different distance measures, which are used to determine how far a data point is from the center of the cluster, could produce various clusters.

5. Data that are categorical cannot be used with the K-Means technique.

6. It's possible that the procedure won't converge in the allotted number of iterations. Convergence should constantly be checked.

## Step 1: Initialization

First, initialize any random points known as the cluster's centroids. When initializing, it's important to remember that the cluster's centroids must be lower than the total amount of training data points. The following two phases are carried out iteratively because this algorithm is an iterative one.

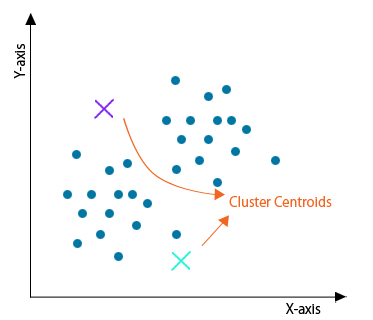


Figure 18 initialize any random points known as the cluster's centroids (Pedamkar, 2022)

## Step 2: Cluster Assignment

The distance between each centroid and each data point is calculated after all data points have been traversed. Now, the minimal distance from the centroids would determine how the clusters were constructed. In this illustration, there are two clusters of the data.

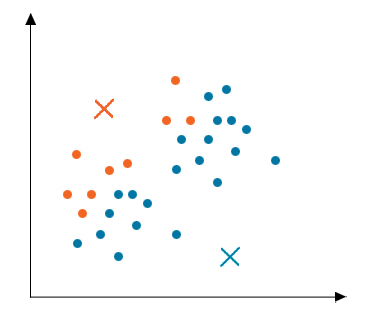


Figure 19 determine the minimal distance from the centroids (Pedamkar, 2022)

## Step 3: Moving Centroid

Since the clusters created in the previous step were not optimized, we must create new clusters that are. To achieve this, we must repeatedly relocate the centroids. Take one cluster's data points, calculate their average, and then relocate the cluster's centroid there. For every other cluster, carry out the same procedure.

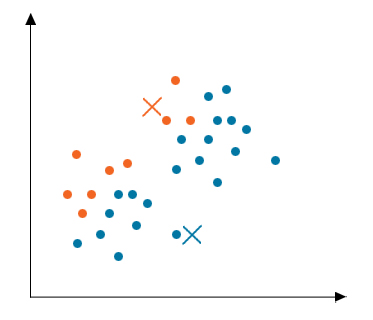


Figure 20 : Moving Centroid (Pedamkar, 2022)

## Step 4: Optimization

## The aforementioned two processes are repeated until the centroids stop moving, that is, until they cease to vary their positions and become static. The k-means algorithm is said to have converged once this is completed.

## Step 5: Convergence

Now that the algorithm has converged, discrete clusters have developed and are readily apparent. Depending on how the clusters were started in the first phase, this technique can produce a variety of solutions.

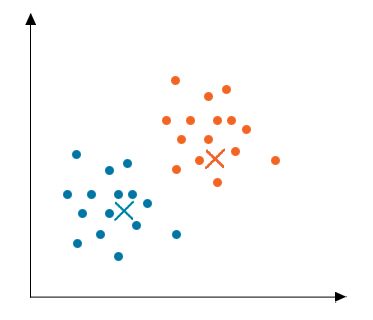


Figure 21 : Convergence (Pedamkar, 2022)

## Importing libraries

import numpy as np *# linear algebra*

import pandas as pd *# data processing, CSV file I/O (e.g. pd.read\_csv)*

import matplotlib.pyplot as plt *# plotting and data visualization*

import seaborn as sns *# improve visuals*

sns.set() *# Set as default style*

import string *# python library*

import re *# regex library*

from gensim.parsing.preprocessing import preprocess\_string, strip\_tags, strip\_punctuation, strip\_multiple\_whitespaces, strip\_numeric, remove\_stopwords, strip\_short *# Preprocesssing*

from gensim.models import Word2Vec *# Word2vec*

from sklearn import cluster *# Kmeans clustering*

from sklearn import metrics *# Metrics for evaluation*

from sklearn.decomposition import PCA *#PCA*

from sklearn.manifold import TSNE *#TSNE*

## Load data

fake = pd.read\_csv('../input/fake-and-real-news-dataset/Fake.csv')

true = pd.read\_csv('../input/fake-and-real-news-dataset/True.csv')

## Data Preprocessing

*Remove tweet disclaimer*

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\_data

true.head(10)

## Merging of fake and True data frame

*# Merging title and text*

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

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

*# Adding fake and true label*

fake['Label'] = 0

true['Label'] = 1

*# We can merge both together since we now have labels*

final\_data = pd.concat([fake, true])

*# Randomize the rows so its all mixed up*

final\_data = final\_data.sample(frac=1).reset\_index(drop=True)

*# Drop columns not needed*

final\_data = final\_data.drop(['title', 'text', 'subject', 'date'], axis = 1)

## Merging of fake and True data frame

*# Here we preprocess the sentences*

def remove\_URL(s):

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

return regex.sub(r'',s)

*# Preprocessing functions to remove lowercase, links, whitespace, tags, numbers, punctuation, strip words*

CUSTOM\_FILTERS = [lambda x: x.lower(), strip\_tags, remove\_URL, strip\_punctuation, strip\_multiple\_whitespaces, strip\_numeric, remove\_stopwords, strip\_short]

*# Here we store the processed sentences and their label*

processed\_data = []

processed\_labels = []

for index, row **in** final\_data.iterrows():

words\_broken\_up = preprocess\_string(row['Sentences'], CUSTOM\_FILTERS)

*# This eliminates any fields that may be blank after preprocessing*

if len(words\_broken\_up) > 0:

processed\_data.append(words\_broken\_up)

processed\_labels.append(row['Label'])

## Word2Vec

*# Word2Vec model trained on processed data*

model = Word2Vec(processed\_data, min\_count=1)

## Sentence Vectors

*# Getting the vector of a sentence based on average of all the word vectors in the sentence*

*# We get the average as this accounts for different sentence lengths*

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** processed\_data:

X.append(Sentence\_Vector(data\_x))

## Clustering Algorithm

## Choice of K

The most important parameter in K means cljustering algorithm is determining the *optimal number of clusters (K)*. There are several methods to find the best value of K. These include the

* Elbow Method
* Silhouette coefficient

#### **Elbow Method**

In this method, a curve is drawn between “within the sum of squares” (WSS) and the number of clusters. The curve plotted resembles a human arm. It is called the elbow method because the point of the elbow in the curve gives us the optimum number of clusters. In the graph or curve, after the elbow point, the value of WSS changes very slowly, so the elbow point must be considered to give the final value of the number of clusters.

**Silhouette coefficient**

The silhouette coefficient is **a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation)**. Select a range of values of k (say 1 to 10). Plot Silhouette coefﬁcient for each value of K.

For this work, the number of clusters is limited to two (2) which represents the target classes of the dataset used in this experiment.

*# Training for 2 clusters (Fake and Real)*

kmeans = cluster.KMeans(n\_clusters=2, verbose=1)

*# Fit predict will return labels*

clustered = kmeans.fit\_predict(X\_np)

correct = 0

incorrect = 0

for index, row **in** testing\_df.iterrows():

if row['Labels'] == row['Prediction']:

correct += 1

else:

incorrect += 1

print("Correctly clustered news: " + str((correct\*100)/(correct+incorrect)) + "%")

**Cluster Visualization**

For the purpose of this research we used both PCA and t-SNE for dimensionality reduction in order to visualize the clusters. A common unsupervised Matrix Factorization technique for reducing dimensionality is PCA. It uses the square covariance (or correlation) matrix to identify a lower-dimensional projection for the original set of variables while retaining a large portion of the original variance in the lower-dimensional covariance matrix. Now, in the new space, each original variable can be projected and represented as a lower-dimensional vector. PCA is frequently used to generate features or visualize them in a lower dimension. Because the embedded space is still high-dimensional and cannot be displayed for the human eye, PCA is frequently advised when visualizing word embeddings, for instance. So, as a bonus, here is the method's link. T-SNE reduces high-dimensional data to a low-dimensional graph, just like PCA (2-D typically). It also works well for dimensionality reduction. t-SNE can reduce dimensions with non-linear relationships, in contrast to PCA. In other words, if the non-linear "Swiss Roll" distribution of our data applied to our data, no change in X or Y would result in a constant change in the other variable.

*# PCA of sentence vectors*

pca = PCA(n\_components=2)

pca\_result = pca.fit\_transform(X\_np)

PCA\_df = pd.DataFrame(pca\_result)

PCA\_df['cluster'] = clustered

PCA\_df.columns = ['x1','x2','cluster']

*# T-SNE*

tsne = TSNE(n\_components=2)

tsne\_result = tsne.fit\_transform(pca\_result)

TSNE\_df = pd.DataFrame(tsne\_result)

TSNE\_df['cluster'] = clustered

TSNE\_df.columns = ['x1','x2','cluster']

For both PCA and T-SNE, all attributes were reduced to two components.

**System requirements**

**Hardware requirements**

The hardware requirements and their specifications of the system for the project implementation.

* PROCESSOR: PENTIUMIV2.6ghz, intelcore2duo
* RAM: 4GB RAM
* LAPTOP HARDDISK: 100GB
* POINTING MOUSE: Optical mouse

**Software requirements**

The software Requirements used for the implementation of the program. Colab Jupyter Notebook

The following libraries are also used;

* Pandas
* Numpy
* Scikit\_learn
* matplotlib
* seaborn

# CHAPTER FOUR

# SIMULATION AND DISCUSSION

## CLUSTER VISULATION

This template aids in finding and displaying data clusters. The K Means technique seeks to cluster your dataset's data points into K different groups. Each observation (data point) will be assigned to the cluster whose center it is closest to once the algorithm has been executed.

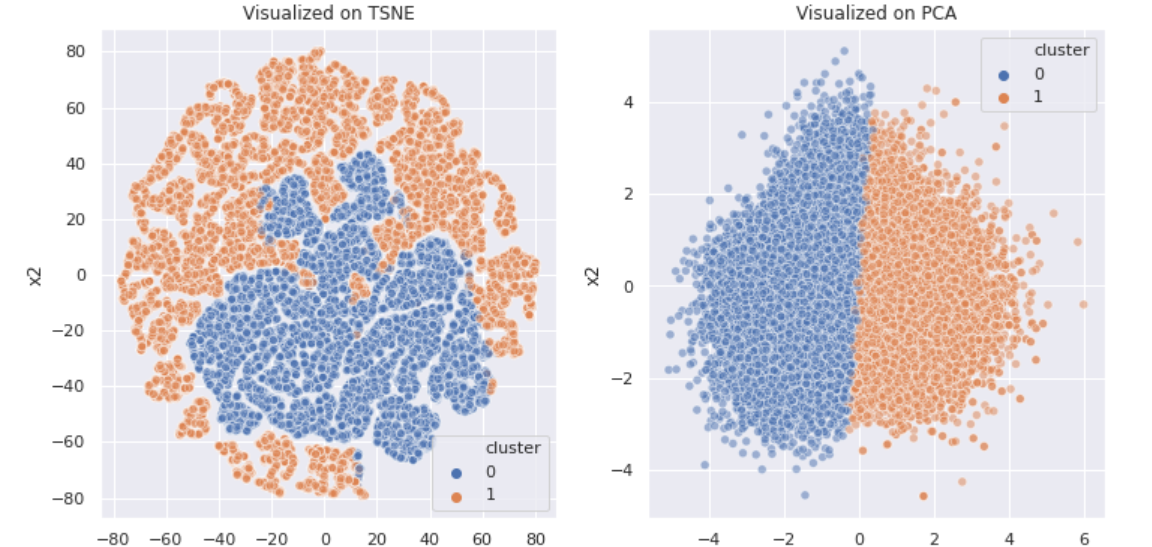


Figure 18 K Means visualization results for 2 components

Here we have managed to clearly segregate the data into subgroups, based on the conditions of K means clustering algorithm. These subgroups have no labels as K means clustering is an unsupervised learning technique but the results can be interpreted to represent the main subject matter of this research; real news and fake news.

## Accuracy

Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used:

In most circumstances, a model with a high accuracy value is a good model, but because we're training a classification model in this case, an item that was predicted is not a good model. False positives can occur when something appears to be true when it isn't.  Similarly, if an article was foreseen, it may have detrimental implications.  This can establish trust because it was marked as fake despite containing genuine material. As a result, we've employed three more metrics, taking into consideration the observation that was erroneously categorized, that is F1-score, precision, and recall.

## DISCUSSION

We tested our model after doing extensive hyperparameter tuning. The model's goal is to be efficient in detecting fake and real and news in a set of datasets.

## ACCURACY

|  |  |
| --- | --- |
| Model | Accuracy |
| K means | 87% |

### Our main measure for this experiment is K means which obtained 87% accuracy in classifying fake. Although other models perform better than K means, the clustering algorithm has an added advantage of relatively simple to implement and also the fact that it can scale to Scales to large data sets.

# CHAPTER FIVE

# SUMMARY, CONCLUSION AND RECOMMENDATION

The work of manually classifying news necessitates a thorough understanding of the domain as well as the ability to spot anomalies in the text. The topic of classifying false news articles using machine learning models and ensemble approaches was tackled in this study. The information we used in our research came from the Internet and consisted of news stories from various domains that covered the majority of the news rather than being expressly classified as political news. The main aim is to investigate the effectiveness of machine learning algorithms to detect fake news.

To achieve optimal accuracy, the learning models were trained and parameter-tuned. Some models have been shown to be more accurate than others. To compare the outcomes of each method, we used a variety of performance indicators. When compared to individual learners, the ensemble learners had a higher overall score on all performance indicators.

There are numerous outstanding issues in the detection of fake news that researchers must address. Identifying essential aspects involved in the distribution of news, for example, is a vital step in reducing the spread of fake news. To identify the primary sources engaged in the dissemination of fake news, graph theory and machine learning approaches can be used. Similarly, real-time fake news detection in videos could be a promising future avenue.

# REFERENCES

A. P. Bradley, R. P.-r.-R.-1. (2006). *Precision-recall operating characteristic (P-ROC) curves in imprecise environments,*. Retrieved from Ieeexplore: https://ieeexplore.ieee.org/abstract/document/1699797

Aldwairi, M., & Alwahedi, A. (2018). Detecting Fake News in Social Media Networks. *The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks*, 215-222.

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 election. *Journal of Economics perspectives*, 211-236.

Allcott, H., & Gentzkow, M. (n.d.). Social Media and fake news in the 2016 election. *Journal of Economic Perspertives, 2*, 211-236.

Bhandari, A. (n.d.). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Bhandari, A. (2020, April 17). *Everything you Should Know about Confusion Matrix for Machine Learning*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Brownlee, J. (2020, August 15). *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Chu, T. (n.d.). *Complete Guide to Parameter Tuning in Gradient Boosting (GBM) in Python*. Retrieved from nttrungmt-wiki: https://sites.google.com/site/nttrungmtwiki/home/it/machine-lear/ensemble-learning/complete-guide-to-parameter-tuning-in-gradient-boosting-gbm-in-python

CKai Shu, S. W. (n.d.). *Exploiting Tri-Relationship for Fake News Detection.*

Club, U. M. (2017). *kaggle*. Retrieved from kaggle: https://www.kaggle.com/c/fake-news/overview

colab, G. (n.d.). *google colab*. Retrieved from google colab: https://colab.research.google.com/drive/1Gl2CB\_-4\_SFtk4RMTVin6tyac8i8rCEv#scrollTo=Gd04QOAl7Owd&uniqifier=2

de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S., & Mattos, D. F. (2021). Identifying Fake News on Social Networks Based on Natural. *Information*.

Dong, X. L. (2014, September ). *From Data Fusion to Knowledge Fusion.* Retrieved from http://www.vldb.org/pvldb/vol7/p881-dong.pdf

H. Ahmed, I. T. (2017). “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques". *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, , pp. 127–138.

H.G. Lewis and M, B. (2010, November 25). *A generalized confusion matrix for assessing area estimates from remotely sensed data*. Retrieved from International Journal of Remote Sensing: https://www.tandfonline.com/doi/abs/10.1080/01431160152558332

Hardman, E. B.-A. (n.d.). Simplifying the visualization of confuion matrix.

Iatropoulou, S. G. (2020, July). Fake News Detection on news articles using Machine and Deep Learning Techniques. *BSc THESIS*. DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION, NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS.

Iftikhar Ahmad, M. Y. (2020, September 16). *Hindawi*. Retrieved from Hindawi: https://www.hindawi.com/journals/complexity/2020/8885861/

Janez, C., & Risius, M. (2017). Automatic Detection of Fake News on Social Media Platforms. *PACIS 2017 Proceddings*, 261.

Junaed Younus Khan, M. T. (2019). A Benchmark Study on Machine Learning Methods for Fake News.

K.P.Singh, D. (2018). *Fake News Detection.*

Kai Shu, A. s. (2017, JUNE). *ACM DIGITAL LIBRARY*. Retrieved from ACM DIGITAL LIBRARY: https://dl.acm.org/doi/abs/10.1145/3137597.3137600

Kim, Y. (2014). “Convolutional Neural Networks for Sentence Classification". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. pp. 1746–1751).

Knoll, A. N. (2013, December 04). *Gradient boosting machines, a tutorial*. Retrieved from frontiers in neurobotics: https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full

Lakshmikumar, P. C. (2019). *Fake News Detection - A Deep Neural Network.*

Leong, L. (. (2017, May 25). *Fighting Fake News: How Google, Facebook and Others Are Trying to Stop It.* Retrieved from Techradar: https://www.techradar.com/news/fighting-fake-news-how-google-facebook-and-more-are-working-to-stop-it

Liu, Y., & Wu, Y.-F. B. (2018). Early Detection of Fake News on Social Media Through Propagation Path. *The Thirty-Second AAAI Conference on Artificial Intellifence(AAAI-18).* New Jersey .

Pedamkar, P. (2022, 7 7). *K- Means Clustering Algorithm*. Retrieved from educba: https://www.educba.com/k-means-clustering-algorithm/

Raschka, S. (2014). *Naive Bayes and Text Classification*. Retrieved from Sebastianraschka: https://sebastianraschka.com/Articles/2014\_naive\_bayes\_1.html

Rastegar-Mojarad, M. &. (2015). *Application of Clinical Text Data for Phenome-Wide Association Studies (PheWASs)*. Retrieved from ResearchGate: https://www.researchgate.net/publication/271910222\_Application\_of\_Clinical\_Text\_Data\_for\_Phenome-Wide\_Association\_Studies\_PheWASs/citation/download

Risdal, M. (n.d.). *kaggle*. Retrieved from kaggle: https://www.kaggle.com/mrisdal/fake-news

Rish, I. (n.d.). An empirical study of the naive Bayes classifier.

S. I. Manzoor, J. S. (2019). *Fake News Detection Using Machine Learning approaches: A systematic Review*. Retrieved from IEEE: https://ieeexplore.ieee.org/abstract/document/8862770

Schmit, M. L. (2015). *ScienceDirect*. Retrieved from Machine Language: https://www.sciencedirect.com/topics/engineering/machine-language

Shu, K., & Liu, H. (2019, July). morgan and claypool publishers. *Synthesis Lectures on Data Mining and Knowledge Discovery, 11*, 1-129. Retrieved from morganclaypool.com: https://www.morganclaypool.com/doi/pdf/10.2200/S00926ED1V01Y201906DMK018

Shu, K., Amy, S., Suhang , W., Tang, J., & Huan, L. (n.d.). Fake News Detection on Social Media:. *SIGKDD Explorations, 19*(1), 22-36.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. 22-36.

Silverman, C. (2016). *This analysis shows how fake election news stories outperformed real news on facebook.* Buzzfeed.

*Simple guide to confusion matrix terminology*. (2014, March 25). Retrieved from Data school: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

*sklearn.metrics.classification\_report*. (n.d.). Retrieved from scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification\_report.html

Stahl, K. (2018). *Fake news detection in social media.*

V, R. (2017). Deception detection and rumor debunking for social media. *HandBook of SOcial Media*.

V. M. Krešňáková, M. S. (2019). Deep learning methods for Fake News detecti. *IEEE*.

V. Rubin, N. C. (2016). “Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News,” . *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, (pp. pp. 7–17).

Wiener, A. L. (2002). Classification and Regression by randon Forest.

Y. Long, Q. L.-R. (2017). Fake News Detection Through Multi-Perspective Speaker Profiles,. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers),*, (pp. pp. 252–256).

Z Khanam, B. N. (2020). Fake News Detection Using Machine Learning Approaches. *IOPSCIENCE*.

Z Khanam, B. N. (2021). Fake News Detection Using Machine Learning Approaches. *IOP Conference Series: Materials Science and Engineering*.

# APPENDIX