FAKE NEWS DETECTION USING MACHINE LEARNING

A PROJECT REPORT

Submitted by

Ankan Das (20BCS5394)

Adars Kumar Singh (20BCS5316)

Swarnim Aditya Pandey (20BCS1313)

Vasu Dhamija (20BCS5279)

Yogesh Kachhawa (20BCS7821)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE & ENGINEERING

Chandigarh University

NOVEMBER & 2022

BONAFIDE CERTIFICATE

Certified that this project report **FAKE NEWS DETECTION USING MACHINE LEARNING** is the bonafide work of **ANKAN DAS, ADARS KUMAR SINGH, VASU DHAMIJA, SWARNIM ADITYA PANDEY AND YOGESH KACHHAWA** who carried out the project work under my/our supervision.

SIGNATURE
Dr. Sandeep Singh Kang
HEAD OF THE DEPARTMENT
Computer Science & Engineering

SIGNATURE
Er. Gazy Abbas Najar
SUPERVISOR
Computer Science & Engineering

Submitted for the project viva-voce examination held on_

INTERNAL EXAMINER

EXTERNAL EXAMINER

TABLE OF CONTENTS

List of Figures

ABSTRACT

With the recent social media boom, the spread of fake news has become a great concern for everybody. It has been used to manipulate public opinions, influence the election - most notably the US Presidential Election of 2016, incite hatred and riots like the genocide of the Rohingya population. A 2018 MIT study found that fake news spreads six times faster on Twitter than real news. The credibility and trust in the news media are at an all-time low. It is becoming increasingly difficult to determine which news is real and which is fake. Various machine learning methods have been used to separate real news from fake ones. In this study, we tried to accomplish that using Logistic Regression, LSTM and natural language processing. There are lots of machine learning models but these two have shown better progress.

Now there is some confusion present in the authenticity of the correctness. But it definitely opens the window for further research. There are some of the aspects that has to be kept in mind considering the fact that fake news detection is not only a simple web interface but also a quite complex thing that includes a lot of backend work.

CHAPTER 1
INTRODUCTION

1.1 Client Identification and Introduction

Fake news detection is a subtask of text classification and is often defined as the task of classifying messages as real or fake. The term "fake news" refers to false or misleading information that looks like real news. It is intended to deceive or mislead people. Fake news is clickbait (misleading headlines), disinformation (malicious to mislead the public), misinformation (false information regardless of motive), hoaxes, parodies, satire, rumors, misleading It appears in various forms, such as news.

Fake news is not a new topic. However, it has become a hot topic since the 2016 US presidential election. Traditionally, people receive their news from trusted sources, media outlets and editors, and usually follow a strict code of conduct. In the late 20th century, the Internet gave rise to new ways of consuming, publishing, and sharing information with little or no editorial standards. Social media has become a major news source for many people these days. According to a Statistical report, footnote 1, there are about 3.6 billion social media users worldwide (about half the population). Social media sites and networks have obvious advantages in disseminating news: B. Immediate access to information, free distribution, no time limits, and diversity. However, these platforms are largely unregulated. Therefore, it is often difficult to determine whether some messages are genuine or fake.

Given a multi-source news dataset and social contexts of news consumers (social media users), the task of fake news detection is to determine if a news item is fake or real. Formally, we define the problem of fake news detection as:

Input: News items, social contexts and associated side information
Output: One of two labels: ‘fake’ or ‘real’.
Fake news can be divided into three groups. The first group is fake news. In other words, News is completely fake and made by the author of article.

The second group is Fake Satire News. This is fake news with the main purpose of providing humour to readers.

The third group is poorly written news articles that contain some real news but are not completely accurate. In short, reporting completely fake stories, using politician quotes for example, is news. These types of messages are typically designed to promote a particular intent or biased opinion [1]. The identified fake news by reviewing the existing literature in his two phases of characterization and detection. investigated the issue of During the characterization stage, they introduced the basic concepts and principles of fake news to both traditional and social media. In the detection phase, we review indexing approaches to detect fake messages from a data mining perspective, such as feature extraction and model building.

The advent of the World Wide Web and the rapid spread of social media platforms (such as Facebook and Twitter) have paved the way for an unprecedented spread of information in human history. With the current use of social media platforms, consumers are creating and sharing more information than ever before, some of which is misleading and irrelevant to reality. Automatically classifying text articles as misinformation or disinformation is a difficult task. Even experts in a given field must examine several aspects before judging the credibility of an article. In this work, we propose using an ensemble machine learning approach for automatic classification of news articles. Our research examines different text properties that can be used to distinguish between fake and genuine content.

We use these properties to train different machine learning algorithms and different ensemble method combinations and evaluate their performance on four real-world datasets. Experimental evaluations confirm the superior performance of the proposed ensemble-learner approach compared with individual learners.

1.2 Identification of Problem

Fake news detection is a subtask of text classification and is often defined as the task of classifying messages as real or fake. The term "fake news" refers to false or misleading information that looks like real news. It is intended to deceive or mislead people. Fake news is clickbait (misleading

headlines), disinformation (malicious to mislead the public), misinformation (false information regardless of motive), hoaxes, parodies, satire, rumors, misleading It appears in various forms, such as news. Fake news is not a new topic. However, it has become a hot topic since the 2016 US presidential election. Traditionally, people receive their news from trusted sources, media outlets and editors, and usually follow a strict code of conduct. In the late 20th century, the Internet gave rise to new ways of consuming, publishing, and sharing information with little or no editorial standards. Social media has become a major news source for many people these days. According to a Statistical report, footnote 1, there are about 3.6 billion social media users worldwide (about half the population). Social media sites and networks have obvious advantages in disseminating news: B. Immediate access to information, free distribution, no time limits, and diversity. However, these platforms are largely unregulated. Therefore, it is often difficult to determine whether some messages are genuine or fake. Social media and the internet are suffering from fake accounts, fake posts and fake news. The intention is often to mislead readers and or manipulate them in purchasing or believing something that isn't real. A system like this would be a contribution in solving a problem to some extent.

Fake news is information that is false or misleading and is presented as real news. The term 'fake news' became mainstream during the 2016 presidential elections in United States. Following this, Google, Twitter, Facebook took steps to combat fake news. However, due to the exponential growth of information in online news portals and social media sites, distinguishing between real and fake news has become difficult.

1.3 Identification of Tasks

Logistic regression Classifier: It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.Timeline. As we are classifying text on the basis of a wide feature set, with a binary output (true/false or true article/fake article), a logistic regression (LR) model is used, since it provides the intuitive equation to classify problems into binary or multiple classes. We performed hyperparameters tuning to get the best result for all individual datasets, while multiple parameters are tested before acquiring the

maximum accuracies from LR model.

1.4 Organization of the Report

This report gives us a brief introduction about the respective project that is Fake news detection using Machine Learning. This report will give a brief introduction to project. The solution to this project is given in this report in brief. Review and problem definition is also part of the next unit.

The design and flow of data models is given in the next chapter. The selection of design and flow is given. Analysis of project is done and explained along with diagrams. The result of the system shows good accuracy in detecting and separating fake news from the real ones.

CHAPTER 2

LITERATURE REVIEW/BACKGROUND STUDY

2.1 Timeline of the reported problem

More and more individuals are choosing to search for and consume news from social media platforms rather than traditional news organisations as we spend a growing amount of our time communicating online through social media platforms. The factors that account for this change in consumption patterns are innate to those social media platforms: I consuming news is frequently less expensive and timelier.

In 2016, 62 percent of U.S. adults reported receiving news via social media, up to only 49 percent in 2012.

The quality of articles on social networks is lower than that of traditional news agencies, despite the advantages that social media offers. However, a lot of faux news, or news pieces containing purposefully incorrect material, is generated online for a variety of reasons, such financial and political benefit, because it's cheap to supply news online and much faster and easier to promote through social media.

According to estimates, the false news incident known as "Pizzagate" is connected to over 1 million tweets "by the conclusion of the presidential contest. Due to the widespread occurrence of this recent phenomena, "fake news "even received the Macquarie Dictionary's 2016 Word of the Year award.Compared to traditional journalism, such as newspapers or television, on social media, it's simpler to share, discuss, and discuss the news with friends or other readers.[1]

For example, during the 2016 U.S. presidential election, it was clear that the most popular fake news was even more widely shared on Facebook than the most widely believed genuine mainstream news. This illustrates how fake news can upset the authenticity equilibrium of the news ecosystem.[2]

Consumers are purposefully led to believe biassed or incorrect information through fake news. For example, some reports indicate that Russia has developed fake accounts and social media bots to

propagate false stories.

In some cases, fake news is simply intended to incite mistrust and confusion among readers, making it difficult for them to tell what is true from what is false and altering how they interpret and react to actual news. to aid in reducing the harmful consequences of fake news (so that both the general public and the news ecosystem can benefit). We must develop tools that can automatically identify bogus news posted on social media.[3]

Even more websites exist that nearly solely manufacture bogus news. Intentionally publishing propaganda, half-truths, disinformation, and hoaxes under the guise of news, they frequently use social media to increase their reach and drive visitors to their websites. The majority of fake news websites' objectives are to change public perceptions of certain issues (mostly political). Examples of similar websites could also be discovered in many other nations, including China, Germany, Ukraine, and the United States of America.[4]

2.2 Proposed solutions

AI and machine learning, according to scientists, could help combat the problem of fake news [5]. There's a good reason for this: recently, AI algorithms have started to perform significantly better on numerous classification issues (image recognition, voice detection, and so on), thanks to cheaper technology and larger datasets.

In their research Mykhailo Granik et al. [6] provide a straightforward method for identifying false news using a naive Bayes classifier. This strategy was put into practise as a software system and evaluated using a collection of Facebook news postings as the test set. They came from three sizable Facebook pages on the right and left, respectively, as well as three sizable websites with mainstream political news (Politico, CNN, ABC News). A classification accuracy of about 74% was attained. Fake news classification accuracy is somewhat worse. This could be due to the dataset's skewness, as just 4.9% of it contains bogus news.

A framework based on several machine learning approaches was provided by Himank Gupta et al. [7] to address a variety of issues, such as the lack of accuracy, BotMaker's temporal lag, and the lengthy processing time required to handle hundreds of tweets in a single second. They started by

gathering 400,000 tweets from the HSpam14 dataset. Then they describe in more detail the 250,000 non-spam tweets and the 150,000 spam tweets. Together with the Top-30 terms that are offering the largest information gain, they also derived several lightweight characteristics. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%

The first unique ML false news detection algorithm was put out by Marco L. Della Vedova et al. [8] and surpasses other methods currently used in the literature by improving accuracy to 78.8%. Second, they used their technique within a Facebook Messenger Chabot and tested it in a real-world setting; they were able to detect bogus news with an accuracy of 81.7%.

The final dataset consists of 15,500 posts from 32 sites (14 conspiracy pages and 18 scientific pages), with more than 2,300,000 likes from more than 900,000 individuals. 6,577 (42.4%) of the postings are non-hoaxes, whereas 8,923 (57.6%) are.

By learning to predict accuracy assessments in two credibility-focused Twitter datasets, PHEME, a dataset of potential rumours in Twitter, and CREDBANK, a crowd-sourced dataset of accuracy assessments for events in Twitter, Cody Buntain et al. [9] develop a method for automating fake news detection on Twitter. They use this technique to analyse tweets that include false news from the BuzzFeed dataset.

They only apply their approach to the collection of well-liked tweets by finding highly retweeted threads of dialogue and classifying tales using the characteristics of these threads.

2.3 Bibliometric analysis

Misinformation, Disinformation, and Mal-information Some surveys focused on three main categories of information disorder: misinformation, disinformation, and mal-information. Misinformation and disinformation have been used interchangeably in much of the discourse on fake news. However, the two categories differ in terms of the degree of falseness and intent to harm. Misinformation is unverified news, but the source/spreader is unaware, and the intention is not to harm the public, while disinformation is unauthentic news to mislead the audience and the source/spreader knows it is false. The third category, mal-information is the deliberate

dissemination of news (which is real) in order to harm a person, specific organization, or country, e.g., leaking private information, or disclosing one's sexual orientation without public interest justification.

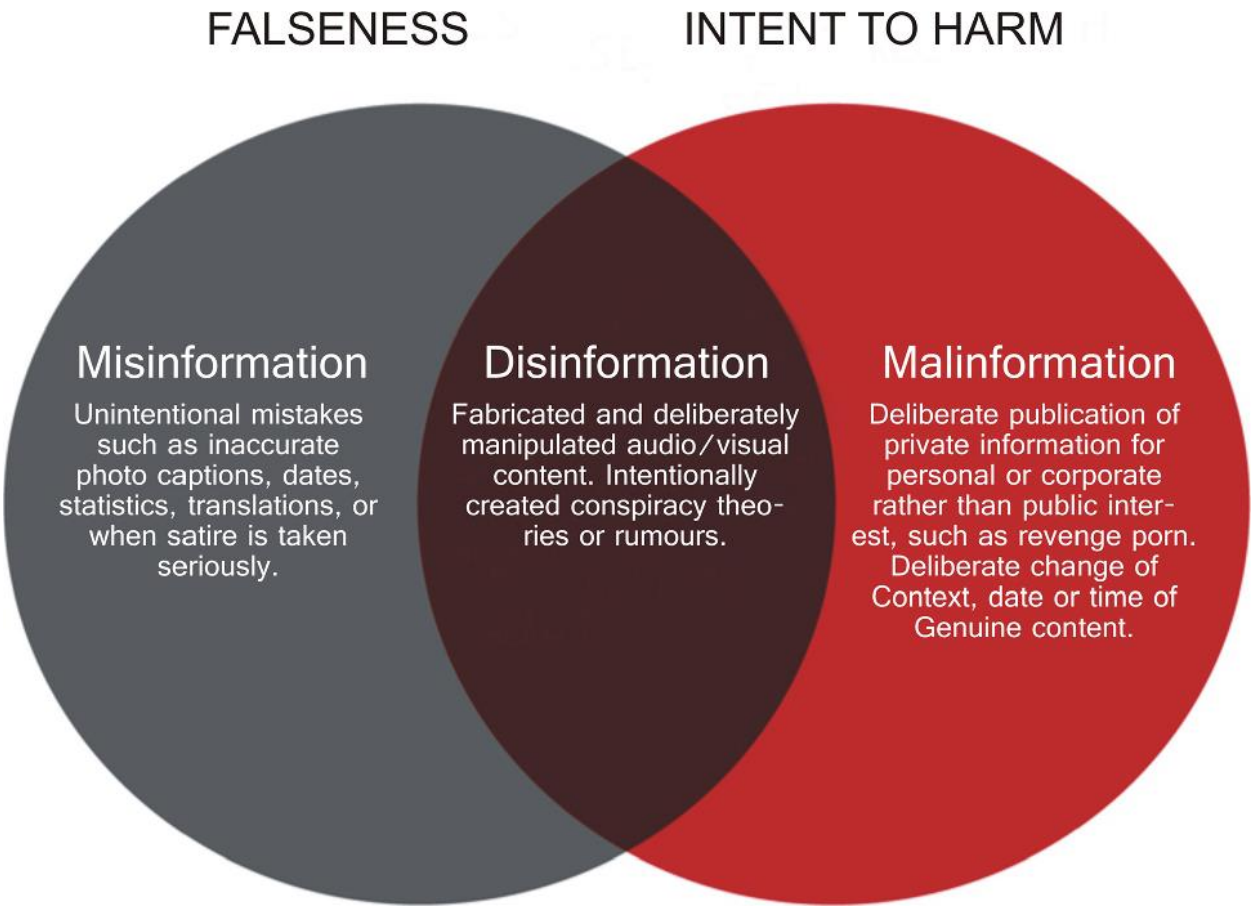

Figure 2.1 Venn diagram for misinformation, disinformation, and mal-information

Majorly, surveys generally classify research based on the methodology. The generic methodological framework is shown in figure 2.2, which shows a roadmap of existing solutions in the literature. In particular, studies have focused on three foremost steps: (1) data collection, (2) feature extraction, and (3) Classification technique. The novelty in research works exists in these steps. Data collection is a challenging task, and researchers have presented different ways to extract data as well as publicly available datasets. Due to the absence of a benchmark dataset for fake news detection, researchers are required to apply different classification models to find the best one suitable for the given dataset. The generic methodology presents different step-by-step paths that have been proposed in the literature. The classification techniques are coarsely grouped into

two categories, machine learning and deep learning-based methods. Machine learning has shown the promising results in this application area of classification. The traditional machine learning approaches are based on handcrafted feature extraction. The generic methodology has shown some handcrafted features, such as linguistic features and propagation patterns, which are prominent in the literature. However, diagnosing relevant features to best capture the deception imposes another challenge. It is time-consuming and may result in biased features, especially in such domains, as fake news detection.

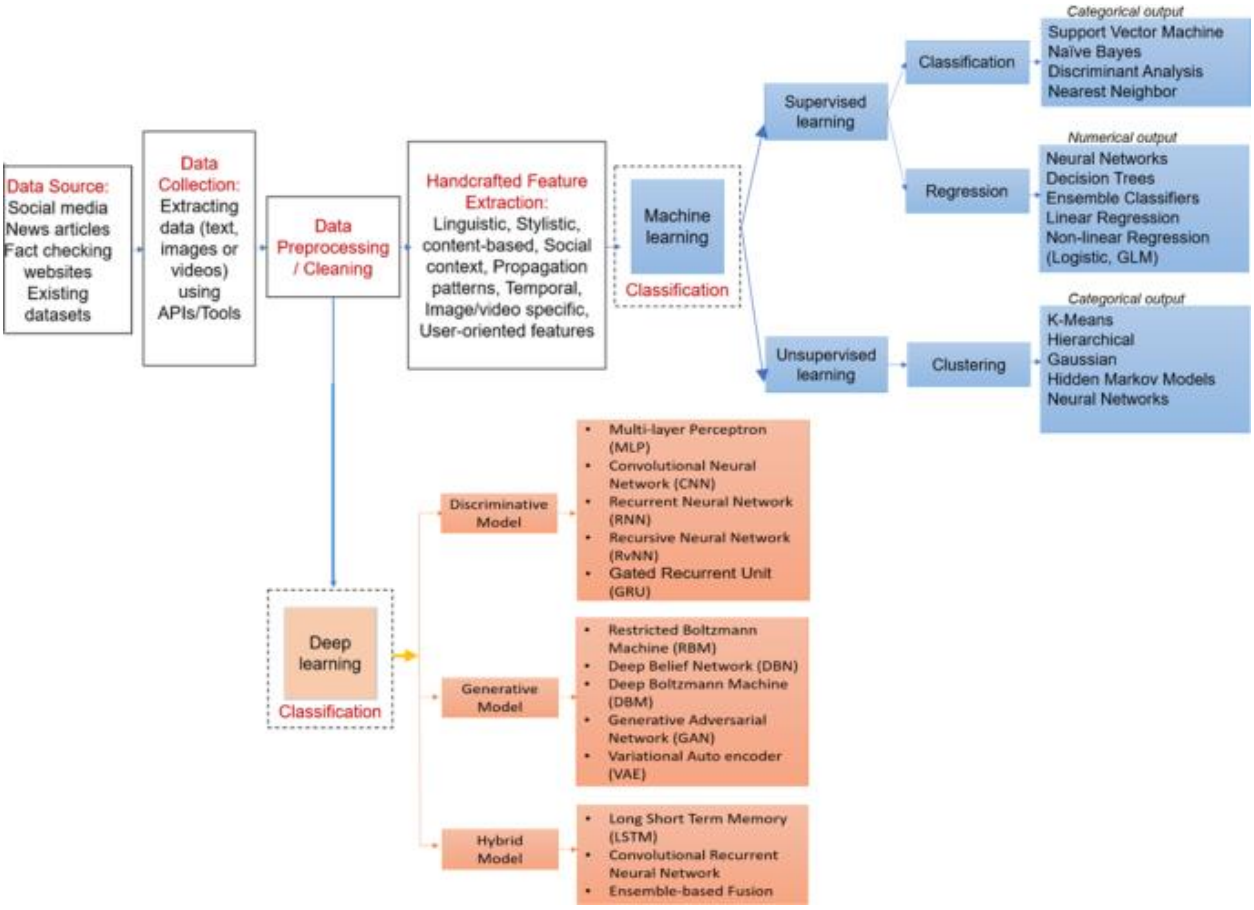

Figure 2.2 Generic methodology for detection of fake news

Perspectives Several surveys reviewed the research based on perspectives. There are four perspectives given the literature for automatic detection of fake news: the unactual knowledge it conveys, its style of writing or content-based, its propagation patterns or social, and the credibility of its source. Figure 2.3 shows the four perspectives with the features used in each. Several research works have been studied in order to describe these perspectives and corresponding

features diagrammatically

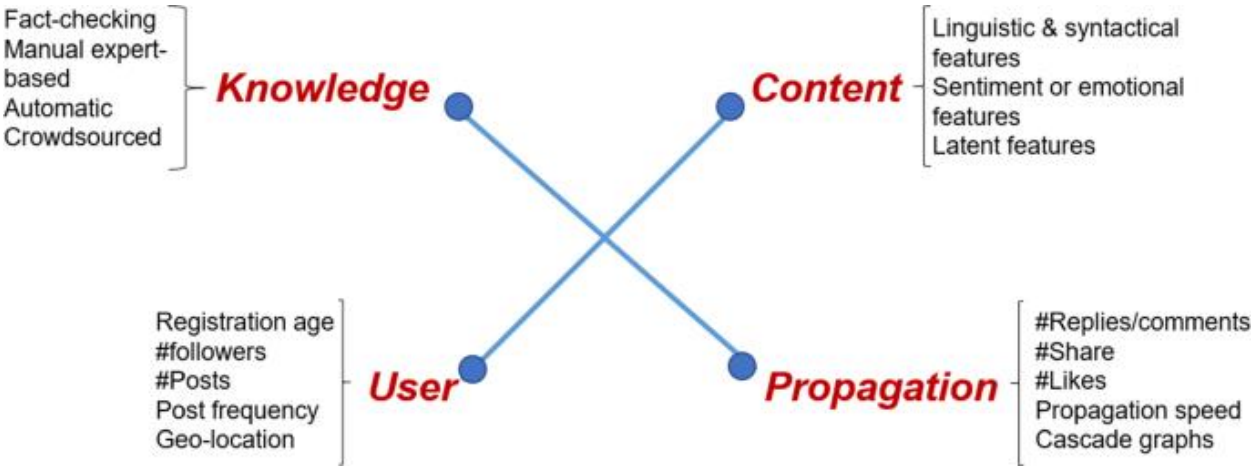

Figure 2.3 Four Perspectives of Fake news detection

2.4 Review Summary

AI and machine learning, according to scientists, could help combat the problem of fake news [10]. There's a good reason for this: recently, AI algorithms have started to perform significantly better on numerous classification issues (image recognition, voice detection, and so on), thanks to cheaper technology and larger datasets.

The system's detection accuracy is somewhere in the range of 70%.

The distinction between this article and others on related topics is that Logistic Regression was used specifically in this paper for the purpose of identifying fake news. Additionally, the developed system was tested on a relatively new data set, giving the opportunity to evaluate how well it performed on recent data.

A number of important articles have been written about automatic deception detection. The authors of give a broad summary of the methods that are currently used to address the issue. The authors discuss their strategy for identifying fake news in [11], which is validated by reader comments for the specific item within microblogs.

2.5 Problem Definition

In our modern era where the internet is ubiquitous, 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 noticeably short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits. In this paper, we aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and check the authenticity of the website publishing the news.

Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, by the development of a system to provide a credible automated index scoring, or rating for credibility of different publishers, and news context. This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

The problem can be broken down into 3 statements:

1) Use NLP to check the authenticity of a news article.

2) If the user has a query about the authenticity of a search query then we he/she can

directly search on our platform and using our custom algorithm we output a confidence score.

3) Check the authenticity of a news source.

These sections have been produced as search fields to take inputs in 3 different forms in our implementation of the problem statement.

The main goal is to apply a set of classification algorithms to obtain a classification model to be used as a scanner for a fake news by details of news detection and embed the model in python application to be used as a discovery for the fake news data. Also, appropriate refactoring has been performed on the Python code to produce an optimized code. The classification algorithms applied in this model are k-Nearest Neighbors (k-NN), Linear Regression, XGBoost, Naive Bayes, Decision Tree, Random Forests and Support Vector Machine (SVM). All these algorithms get as accurate as possible. Where reliable from the combination of the average of them and compare them.

2.6 Goals/Objectives

The fake news on social media and various other media is wide spreading and is a matter of critical concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it. This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tiff Vectorizer. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.The results of the analysis of the datasets using the six algorithms have been depicted using the confusion matrix.

The research in this paper focuses on detecting the fake news by reviewing it in two stages:

characterization and disclosure. In the first stage, the basic concepts and principles of fake news are highlighted in social media. During the discovery stage, the current methods are reviewed for detection of fake news using different supervised learning algorithms.

In the current fake news corpus, there are several examples of text classification using both supervised and unsupervised learning algorithms. However, most of the literature focuses on specific datasets or domains, most notably Policy His domain. Therefore, the trained algorithm works best in a particular article type domain and does not give optimal results when exposed to articles in other domains. Articles from different domains have unique text structures, making it difficult to train a general algorithm that works best for all specific news domains. In this article, we propose a solution to the problem of fake news detection using an ensemble machine learning approach. Our research examines various text properties that can be used to distinguish between fake and genuine content. By leveraging these properties, we train different combinations of machine learning algorithms and different ensemble methods, which have not been fully explored in the current literature. Ensemble learners have proven useful in a variety of applications because learning models tend to use techniques such as bagging and boosting to reduce error rates.

These techniques facilitate training of various machine learning algorithms in an effective and efficient manner. We also conducted large-scale experiments on four publicly available real-world datasets. The results confirm the performance improvement of the proposed method using four commonly used performance metrics (that is, accuracy, precision, recall, and F-1 score).

Fake news detection is a subtask of text classification and is often defined as the task of classifying messages as real or fake. The term "fake news" refers to false or misleading information that looks like real news. It is intended to deceive or mislead people. Fake news includes clickbait (misleading headlines), disinformation (malicious things that mislead the public), misinformation (false information regardless of motive), hoaxes, parodies, satire, rumors, and misunderstandings. There are various forms such as inviting news. discussed in the literature.

CHAPTER 3
DESIGN FLOW/PROCESS

3.1 Evaluation & Selection of Specifications/Features

Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages.

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words “chocolates”, “chocolatey”, and “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”.

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

3.2 Design Constraints

From an NLP perspective, researchers have studied numerous aspects of the credibility of online information. For example, applied the time-sensitive supervised approach by relying on tweet content to address the credibility of a tweet in different situations. We used LSTM in a similar problem of early rumour detection. In another work, aimed at detecting the stance of tweets and determining the veracity of the given rumour with convolution neural networks. A submission to the SemEval 2016 Twitter Stance Detection task focuses on creating a bag-of-words autoencoder

and training it over the tokenized tweets. Another team, [13], combined multiple models in an ensemble providing a 50/50 weighted average between a deep convolutional neural network and a gradient-boosted decision tree. Though this work seems to be similar to our work, the difference lies in the construction of an ensemble of classifiers.

In a similar attempt, a team [14] concatenated various features vectors and passed them through an NLP model. Passive Aggressive algorithm is a margin-based online learning algorithm for binary classification. It is also an algorithm of a soft margin-based method and robust to noise. It can be used in fake news detection Term Frequency-Inverse Document Frequency is also a method used to represent text in Fake News Detector 12 a format that can be easily processed by machine learning algorithms. It is a numerical statistic that shows how important a word is to news in a news dataset. The importance of a word is proportional to the number of times the word appears in the news (fake and real) but inversely proportional to the number of times the word appears in the news dataset (fake or real)

3.3 Design Flow

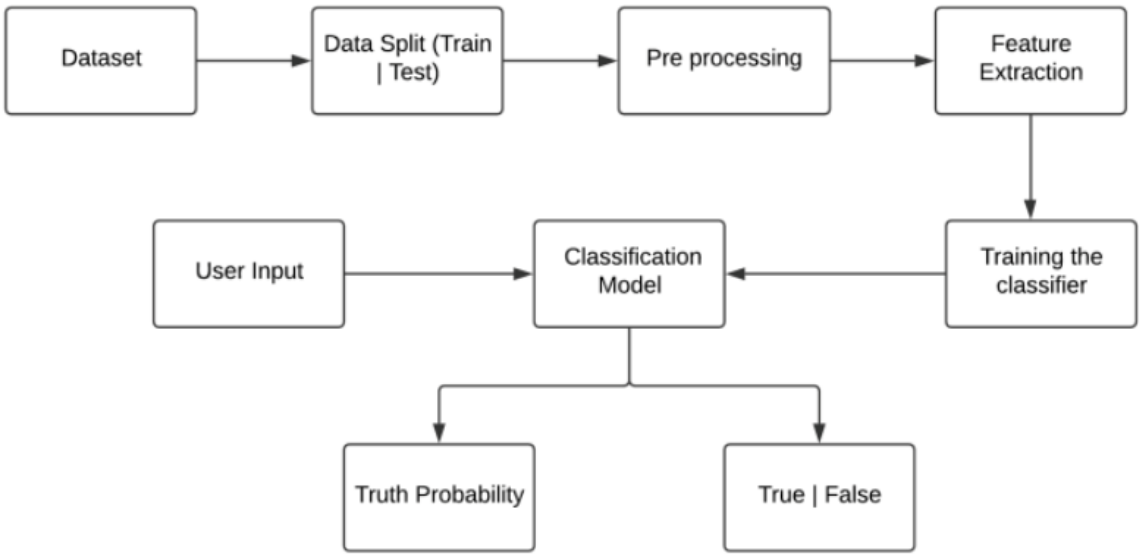

Figure 3.1 Design Flow Diagram

3.4 Design selection

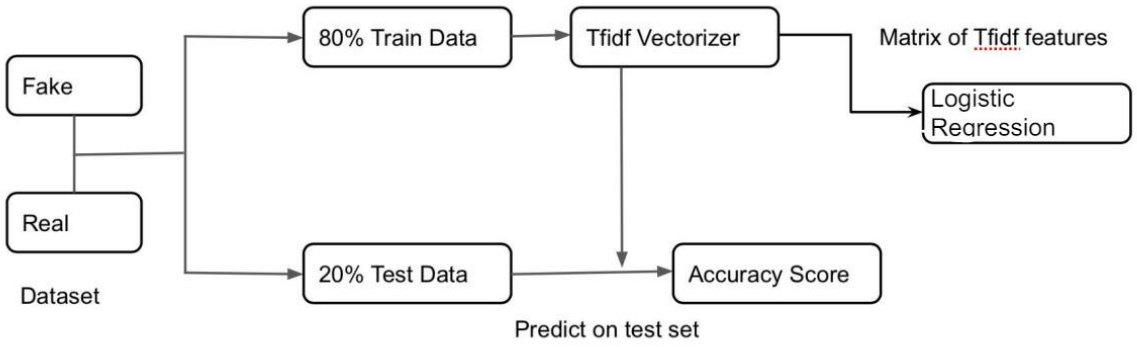

Figure 3.2 Design Selection Diagram

Our target value is of binomial class, for these type of classes, logistic regression yields the best result with high accuracy.

Logistic Function (Sigmoid Function):

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1.

The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

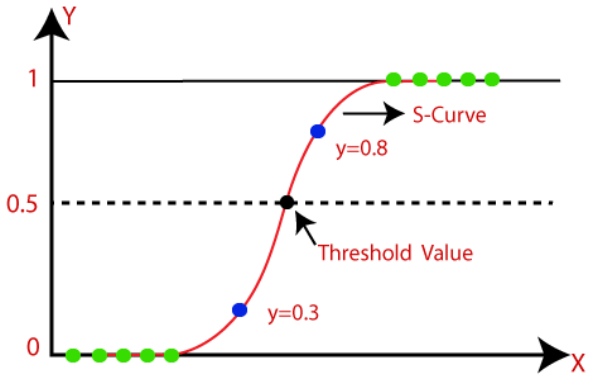

Figure 3.3 Sigmoid Function

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

3.5 Implementation plan/methodology

There are two parts to the ML Model building. Machine Learning is a part of our life that can help us in predicting. We are using two types of models in this case. For the first part, we used passive-aggressive classifiers. And the steps include:

1. Data Loading: We are loading a CSV file for the data sorting and training-testing part of the model. The CSV file is turned into an array for easier work purpose.

2. Vectorization: Vectorization is needed for determining the frequency of the words present in a passage. This is needed to determine which words are used often.

3. Logistic Regression: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

4. Model Building: The model is built through the train and test of the dataset, by ensuring that the training is done for 80% of the dataset and testing is done in the rest of the 20% of the dataset.

In the second part, we used to be LSTM. Here are the steps:

1. Loading the data: For this step, it is the same as the passive-aggressive

one.

2.Scanning and parsing. Data is loaded from a CSV file. This consists of the body of selected news articles. It then contains a label field that indicates whether the news is real or fake. In this code block, we scan the CSV and clean the titles to filter out stop words and punctuation.

3. Tokenization. The tokenizer is used to assign indices to words, and filter out infrequent words. This allows us to generate sequences for our training and testing data.

4. Embedding matrix: Apply the embedding matrix. An embedding matrix is used to extract the semantic information from the words in each title.

5. Model Building: Building the model and finding out the accuracy via confusion matrix. The model is created using an Embedding layer, LSTM, Dropout, and Dense layers. We are going to run the data on 20 epochs.

We observed that the LSTM model is vastly inaccurate in predicting the authenticity of the news. So, we decided to show the output by running it through the Logistic Regression model.

CHAPTER 4

RESULTS ANALYSIS AND VALIDATION

4.1 Implementation of solution

train.csv: A full training dataset with the following attributes:

id: unique id for a news article

title: the title of a news article

author: author of the news article

text: the text of the article; could be incomplete

label: a label that marks the article as potentially unreliable

1: fake news

0: real news

Using Logistic Regression Model because this is Binary Classification system, i.e., real or fake.

train.csv:

id	title	author	text	label
0	House Dem Aide: We Didn���t Even See Comey���s Letter Until Jaso	Darrell Lucus	House Dem Aide: We Didn���t Even See	1
1	FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart	Daniel J. Flynn	Ever get the feeling your life circles the roundabc	0
2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29,	1
3	15 Civilians Killed In Single US Airstrike Have Been Identified	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstrike	1
4	Iranian woman jailed for fictional unpublished story about woman stoned t	Howard Portnoy	Print	1
5	Jackie Mason: Hollywood Would Love Trump if He Bombed North Korea ove	Daniel Nussbaum	In these trying times, Jackie Mason is the Voice o	0
6	Life: Life Of Luxury: Elton John���s 6 Favorite Shark Pictures To Stare At	nan	Ever wonder how Britain���s most iconic	1
7	Beno��t Hamon Wins French Socialist Party���s Presidential Nominat	Alissa J. Rubin	PARIS ��� France chose an idealistic, traditic	0
8	Excerpts From a Draft Script for Donald Trump���s Q&ampA With a Bla	nan	Donald J. Trump is scheduled to make a highly ar	0
9	A Back-Channel Plan for Ukraine and Russia, Courtesy of Trump Associates	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as natic	0
10	Obama���s Organizing for Action Partners with Soros-Linked ���Indi	Aaron Klein	Organizing for Action, the activist group that mor	0
11	BBC Comedy Sketch "Real Housewives of ISIS" Causes Outrage	Chris Tomlinson	The BBC produced spoof on the ���Real Hou	0
12	Russian Researchers Discover Secret Nazi Military Base ���Treasure Hun	Amando Flavio	The mystery surrounding The Third Reich and	1
13	US Officials See No Link Between Trump and Russia	Jason Ditz	Clinton Campaign Demands FBI Affirm Trump's	1
14	Re: Yes, There Are Paid Government Trolls On Social Media, Blogs, Forums A	AnotherAnnie	Yes, There Are Paid Government Trolls On Social	
BART SIMPSONSON				
Hey	it���s just another means of getting the prog���s agitprop out there	channels	and programs fellating them daily�������. James	
It���s r	I imagine most governments do it. And it���s not just governments eith	oil companies spreading disinformatior	difficult to know who to trust on the Internet these days. We all seek out tl	
In any soci	most people do nothing. It���s up to the minority to defend the naive majority. It���s how things are done. Bob G			
If I read the article correctly the government is targeting conservative thought. I always wondered why liberals would deliberately read conservative web sites and then harass the commentators. I certair				
The DNC is	stupid and racist. (Not to say that there are not elements of that in many cc	but these j@ck@sses ramp it up to 11.) Tami Chapman		

Sample Code

```
import numpy as np
import pandas as pd
import re #regular expression library, useful for searching text
from nltk.corpus import stopwords #text that doesn't add value
from nltk.stem.porter import PorterStemmer #returns root words, stemming words
from sklearn.feature_extraction.text import TfidfVectorizer #text into feature number
from sklearn.model_selection import train_test_split #splitting data into train and test
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

```
import nltk
nltk.download('stopwords')
```

Data preprocessing

```
# replaceing null values with empty string
news_dataset = news_dataset.fillna('')
```

ource of news is very effective,author and title would yield good result. Including text, which are very large paras, would ease processing time drastically, giving performance hit.

```
# merging author name and news title
news_dataset['content'] = news_dataset['author'] + ' ' + news_dataset['title']
```

```
port_stem = PorterStemmer()
```

```
def stemming(content):
  stemmed_content = re.sub('[^a-zA-Z]', ' ', content) # only includes alphabets, and replace any ot
  stemmed_content = stemmed_content.lower() # convert all words to lowercase
  stemmed_content = stemmed_content.split() # convert all words to list
  # removing stopwords and converting every word to its root word
  stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.word
  stemmed_content = ' '.join(stemmed_content)
  return stemmed_content
```

```
news_dataset['content'] = news_dataset['content'].apply(stemming)
```

```
# converting the textual data to numerical data
vectorizer = TfidfVectorizer() # gives values to words according to its frequency, into feature vec
vectorizer.fit(X)
X = vectorizer.transform(X)
```

Training the Model: Logistic Regression

```
model = LogisticRegression()
```

```
model.fit(X_train, Y_train)
```

```
LogisticRegression()
```

```
# accuracy score on training dataset
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
print('Accuracy score of the training data: ', training_data_accuracy)
```

```
Accuracy score of the training data:  0.9865985576923076
```

4.2 Result and Analysis

Our model is 98.6% accurate on training data.

```
# accuracy score on training dataset
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
print('Accuracy score of the training data: ', training_data_accuracy)
```

```
Accuracy score of the training data:  0.9865985576923076
```

Our model is 98.6% accurate on training data

```
# accuracy score on test dataset
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
print('Accuracy score of the test data: ', test_data_accuracy)
```

```
Accuracy score of the test data:  0.9790865384615385
```

Our model is 97.9% accurate on testing data

Making a Predictive System

```
X_new = X_test[878]
prediction = model.predict(X_new)
print(prediction)
if (prediction[0] == 0):
  print('News is real')
else:
  print('News is fake')
```

```
[0]
News is real
```

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this paper, we've used Logistic Regression classifier which will serve the model and work with the user input. Here, we've presented a detection model for fake news using analysis through the lenses of different feature extraction techniques. We have investigated different feature extraction and machine learning techniques. The proposed model achieves accuracy of approximately 98.6% on training dataset and 97.9 on testing dataset using logistic regression classifier.

The task of manually classifying messages requires detailed technical knowledge and expertise to identify anomalies in text. In this study, we discussed the problem of classifying fake news articles using machine learning models and ensemble techniques. The data we used in our work comes from the World Wide Web and contains news articles from various fields to cover most news rather than categorizing it specifically for political news. The main purpose of this study is to identify textual patterns that distinguish fake articles from genuine news. We used the regression to extract various text functions from the articles and used the function set as input to the model. The learning models were trained and parameter-tuned to obtain optimal accuracy. The ensemble learners have shown an overall better score on all performance metrics as compared to the individual learners.

Fake news detection has many open issues that require attention of researchers. For example, identifying the key factors responsible for the spread of news is an important step in curbing the spread of fake news. Graph theory and machine learning techniques can be used to identify key sources responsible for the spread of fake news. Similarly, identifying fake news in videos in real time could also be a future direction.

5.2 Future work

Further this project work can be enhanced by using all of this logic work in real time where this is used on social media websites and it shows if the news is fake or not. This will help people from believing true events and not trust fake news.

Due to concept drift, the model trained on our datasets may have biases , causing some legitimate news sites to be incorrectly labelled. This may necessitate a re-labelling and re-evaluation process using more recent data.

We also observe that the sequence length depends on the average sequence length of the dataset. Since our datasets are large and by default contain longer sequences, we get better performance with a larger sequence length. Due to the resource limitations, we could not test on further larger lengths, which we leave for future work.

The producers of disinformation change their tactics over time. We also want to see how these tactics evolve and incorporate these changes into our detection models.

We evaluate our models on a binary classification problem. Our next step will be to consider multi-label classification, which will broaden the model's applicability to various levels of fake news detection.

User connections information can reveal the social group to which a user belongs and how the social network nurtures the spread of fake news. Second, we may incorporate user historical data to better estimate the user status, as a user's tendency to spread fake news may change over time.

REFERENCES

[1] Great moon hoax. https://en.wikipedia.org/wiki/Great Moon Hoax. [Online; accessed 25-September-2017].

[2] S. Adali, T. Liu, and M. Magdon-Ismail. Optimal link bombs are uncoordinated. In AIRWeb, 2005.

[3] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. In ICWSM, 2013.

[4] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 2017.

[5] E. Arisoy, T. Sainath, B. Kingsbury, and B. Ramabhadran. Deep neural network language models. In WLM, 2012.

[6] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In SIGIR, 1998.

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

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

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

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

[11] Mykhailo Granik, Volodymyr Mesyura, “Fake News Detection Using Naive Bayes Classifier”, 2017 IEEEFirst Ukraine Conference on Electrical and Computer Engineering (UKRCON).

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

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

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