FAKE NEWS DETECTION USING DATA MINING TECHNIQUES

BY

MD: Mehedi Hasan
ID: 161-15-925

MD: Harun-Or-Rashid
ID: 161-15-844

Shormy Islam
ID: 161-15-835

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Dr. S. M. Aminul Haque
Associate Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2019

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Sheikh Abujar, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

Supervised by:

_____

Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Dr. S. M. Aminul Haque
Associate Head
Department of CSE
Daffodil International University

Submitted by:

_____

Md. Mehedi Hasan
ID: 161-15-925
Department of CSE
Daffodil International University

_____

Md. Harun-Or-Rashid
ID: 161-15-844
Department of CSE
Daffodil International University

_____

Shormy Islam
ID: 161-15-835
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

We really grateful and wish our profound our indebtedness to **Sheikh Abujar**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining and Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Syed Akhter Hossain, Head, Department of CSE,** for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Social Media is becoming the most popular web site to seek news day by day because of the easy access facility worldwide. It's very cost-effective and people can easily collect news & entertainment from any corner of the world with just a simple click. It's helping the world to be open on the other hand it's true that a rumor can make disaster within a minute which is very easy to spread by such open media. The availability of low-quality news and false information can mislead the readers & which is done intentionally by a group of people. In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources. They intend to spread rumor & earn revenue by making advertisements on their sites. This could make sufferers a large number of peoples at a time. In this paper, we focus on the automatic identification of fake news by using a novel algorithm that's "decision tree algorithm". We may not stop fake news from being made but we can limit to share it. To make limitations on any site, we need assistance from the concerned department of a state or government. Our target is to select headlines of news & send them to the algorithm as well as stop the spread of news which is identified as fake news by the decision tree algorithm. To be successful, we need the help of the central information cell of a country. Our vision is to stop the deceptive information & rumor by limiting the propagation of fake news in social media as well as web sites. It is very challenging but our novel algorithm will perform well to detect fake news and able to get high accuracy over time.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1
INTRODUCTION

1.1 Introduction

Communication plays a vital role in moving the modern world forward. Now this communication system has become much easier. Now almost all of us are leaning on the news online. But sadly, the fact is, many news portals are now spreading fake news. Again some people have been relying on these fake news stories. This has led to many adverse effects in various sectors of society. Political sports business entertainment is also having an adverse effect in many sectors for all these fake news, and people are facing harassment in various ways. Which is never good for us. Now is the time to detect fake news by showing the proportion of wrong news among people.

Every minute, many news posts are published from various news protocols on the revolutionary web of the world, say CNN, BBC, BuzzFeed, Daily star, PolitiFact, etc. [1]. The main intention of this paper is to automatically detect the fake news talk. We collect some datasheets from real-life and make the news content relevant to the news through knowledge-based context and style best method and by identifying and analyzing it [1].

Based on the features of this fake news we develop an accuracy set, in the end, our accuracy rate stands at 82% [3].

1.2 Motivation of research

One of the most important tasks of detecting fake news is to classify the news. Then take these two news stories classified identifiers and analyze their properties. Before that, our biggest motivation is what some news people are looking for more and more, we make those news headlines and perform their work by creating their database. When we have worked with all this news, we have found the news in various categories which are in high demand in the context of our Bangladesh, such as sports, politics and commercial news, etc. All this news is mainly in Bangladesh with a lot of demand from others. When

we collect all this news we are only able to identify. It has helped us a lot by temporarily verifying which news people are most likely to find.

If the trend of human harassment happens for these fake news, then our social world will undoubtedly turn to bad. So from here on out, we have to identify and classify news by identifying their properties [2]. In our research, we spread important news to people by identifying fake news or accurate news, which tends to be 82% accurate [4].

1.3 Rationale of the Study

This research paper will pave the way for the advancement of our personal lives and society. Although this is a challenging one, we can say that fake news detection has a lot to do with it. We collect relevant articles from different journals. And from there we can learn about algorithms for classifying news. We also get ideas from various web articles and blogs, how to detect all these fake news stories through some algorithm.

1.4 Research Questions

- Can we collect row data or noisy data from the daily newspapers or social media?
- Can we pre-process the noisy data to be used for the machine learning approaches?
- Can we find frequent patterns, correlations using association rule in data mining?
- Can we have classified the data after preprocessing data?
- Can we properly detect or identify the category of the given dataset by using machine learning approaches?

1.5 Expected Output

It’s very challenging for users to detect fake news from posted news social media & it’s very important to infer deceptive information to prevent the rumor. As fake news can spread by the author without any obstacles on social media & our research will help to limit this by using our algorithm. We all know that it’s difficult to stop the propagation of rumors, but we can prevent it by limiting the share. Our qualitative & quantitative

datasets can cover the fake topic and able to detect fake news within a very short time. If we can minimize the spreading of fake news, then rumors will not spread vastly.

Already the rumors & fake news are the most disastrous word days & this will affect our next generation.

Hope by our model we may get better output & high accuracy when detecting any fake news.

1.6 Report Layout

The report will be followed given by instructions:

Chapter 1, provides a summary of this research. Introduction and discussion is the key point of this first chapter. The motivated part explained well in this chapter. The most important thing is the rational study also included here, finally what will be the research questions and what will be expected the outcome of this research that has explained in the last section.

Chapter 2, what is already done before researches this topic? How their objectives and what were is the exact goal in this research? What are the problems and what will be solutions in this research are described? In the last part, the research layout has given.

Chapter 3, in this section theoretical discussion about research, has given and the statistical methods of this work are given. This chapter has been shown the procedural approaches of the machine learning classifier, and the last part has described to validation the model also confusion matrix analysis is being presented.

Chapter 4, what are the exact outcome of this research that explained. Some research related pictures have given to understand easily what the criteria of this work are.

Chapter 5, Conclusion part has given in this part. It's very important to the section. A whole research report in the nutshell has been explained. And what are the limitations to do this research, that can be benefited in future researchers, this kind of job?

CHAPTER 2
BACKGROUND

2.1 Introduction

We are this portion of our works by a short introduction as below;

Related works where we'll discuss the researches done by the previous researcher about Fake News Detection. This is inspired by different web-based journals. Research Summary is a short brief about our work where we tried to frame our initiatives by a simple story. The scope of problems is the major place where we'll face troubles during our research. Challenges are the technical aspects where we face difficulties.

2.2 Related Works

A lot of research done based on fake news detection & deceptive information. We're pointing out and summarize some of these previous Related Works following some perspectives that are on below;

Fake news detection methods generally focus on using news contents and social contexts (Shu et al. 2017). The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing.

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing.

2.2.1 Detection Based on Content

The headline of the news is indicating the content of any news. The headline is a mini sentence but represents the whole information of news within a second in front of the reader. A large number of people go through the detailed news based on the headline. It's very easy to detect deceptive information by filtering the content. The working procedure will be based on extracting different headlines to detect fake news by using the traditional Decision Tree algorithm [5]. Some researchers have investigated in the field of detection of fake news, the fake information and achieved quality results.

Nowadays the rumor & fake news is vastly spreading by using different images with the headline of a link in social media. These pictures express some important information & readers got attracted by these images. Visual impressions are essential phenomena for fake news propagation. Fake news producers often use fake images or videos to manipulate the emotions of readers. Gupta A, Lamba H, Kumaraguru P, Joshi A. publish research based on a prediction of the fake image from a real image by using the Decision Tree Classifier. They achieved desirable results [8].

2.2.2 Detection based on Context

The main information of news is presenting by the heading because it is easily visible but the context of news is not easily visible. We need to utilize user social engagements & feedback on this news to detect the deceptive information which is intended to be spread. The context of news usually includes the author details, publisher details, comment on the news article and many more. After analyzing the origination and the circumstance of the news article that spread on social networks, the authenticity of it can be controlled [6]. Each user has the right to represent its unique, in the social network, the profile, content, and others constitute its unique features.

2.2.3 Unmasking Style Categories

. In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for

readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources.

2.2.4 Style Features and Feature Selection

Our writing style model incorporates commonly used style features as well as some specific to the news domain. . In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources.

2.2.5 Baselines

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing.

2.2.6 Hyper partisanship vs. Mainstream

Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing. A. Unmasking hyper partisanship. The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On

each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing.

In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources.

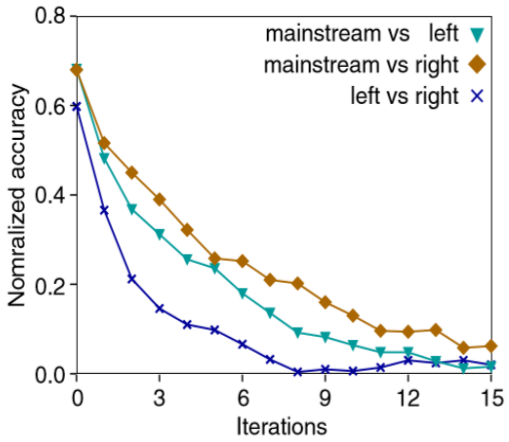


Figure 2.2.6.1: Unmasking applied to pairs of political orientations. The quicker a curve decreases, the more similar the respective styles are.

.In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek

the attention of readers which made it challenging to identify reliable and authorized news sources.

2.3 Research Summary

In his paper, we see to the automatic detection of fake news by using a novel algorithm that's "decision tree algorithm". We may not stop fake news from being made, but we can limit to share it [10]. To make limitations on any site, we need assistance from the concerned department of a state or government. Our target is to select a small subset of news then send them to the algorithm as well as stop the spread of news which is identified as fake news by the decision tree algorithm & help of central information cell of a country. The main focus of our work is to minimize the spread of misinformation by stopping the propagation of fake news in social media. The main challenge in this research is collecting quality data, i.e. instances of fake and real news articles on a balanced distribution of topics. It is very challenging to achieve this objective but our novel algorithm performs well to detect fake news and able to get high accuracy over time.

2.4 Scope of the Problem

As we have to run with a big dataset & it's really difficult to collect a big amount of data as well as check the competencies with the model. Language pattern detection is also a vital difficulty for fake news detection [6]. Performing automatic text classification to differentiate between fake news & real news. We determined to solve this problem with our efforts but the users of different social media might not show interest to share their opinion through our model because they are not abiding to share their thoughts with us. In that case, we have to make a social awareness regarding the issue. Social awareness creation will be a scope of the problem for us.

2.5 Challenges

Our model is based on a quality dataset & we have to face below challenges in the long run.

- Collecting a lot of quality datasets.
- Data cleaning & authenticating data sources.
- Data preprocessing.
- Datasets preparation & published for research purposes.
- Prioritize of preset data & collective data.
- Decision making on tested data.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, we discuss the theoretical knowledge of our research. We know that working with some algorithms and datasets will make the job easier and more accurate. Even this is enough to give a clear idea of the job. In this chapter, we will discuss what instrument and technology we will use for research. Data Collection is how we analyze it through machine learning or data mining, and we finally implement it in a way that is best for all of us

.

3.2 Research Subject and Instrumentation

The main thing about research is the research subject and what area it is working on. The main focus of our research is news-centric. The main thing about Research is the Research subject and what area it is working on. And we take the news and present it through special processing. It is not enough to give a clear idea of research. So in order to have a clear idea of the research in its entirety, you need to know about its fields [7]. We have discussed all the fields for the benefit of this research. There are many types of instrumentations used to construct a research paper that researchers use to facilitate research.

3.3 Data Collection Procedure

One of the main points of every research is data and without this data, it is almost impossible to research. Research is used for various test purposes. And with this data, we can move forward with research. Our research data is based on social news. Social news is a big part of our research. For our research, we collect some past news from our country's most popular English news Daily Star [1]

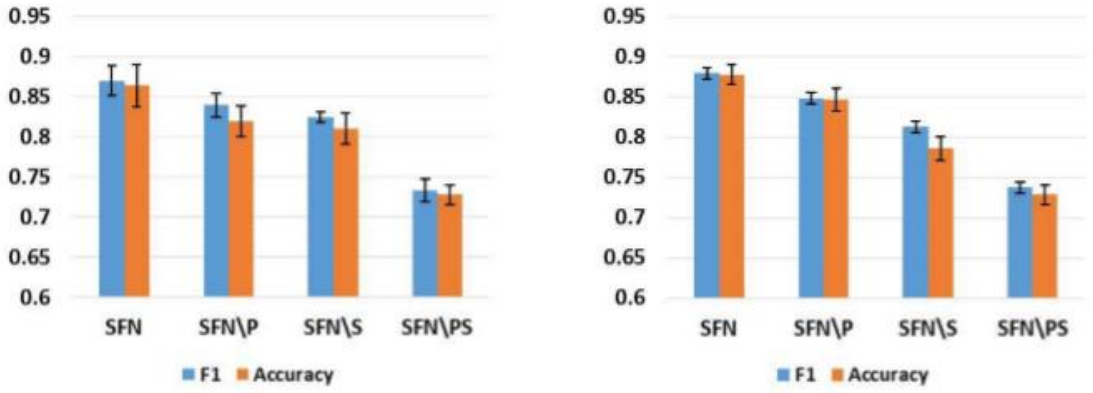
3.4 Statistical Analysis

In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources.

Platform	**BuzzFeed**	**PolitiFact**
# Candidate news	182	240
# True news	91	120
# Fake news	91	120
# Users	15,257	23,865
# Engagements	25,240	37,259
# Social Links	634,750	574,744
# Publisher	9	91

Table 3.4.1: The statistics of datasets

The publishers' partisan labels are collected from a well-known media bias fact-checking websites MBFC. Note that we balance the number of fake news and true news, so that we avoid that trivial solution (e.g., classifying all news as the major class labels) to achieve high performance and for fair performance comparison. The details are shown in Table 3.4.1.



(a) BuzzFeed (b) PolitiFact

Figure 3.4.2: User engagements and publisher partisan impact.

The parameters in all the variants are determined with cross-validation and the performance comparison is shown in Figure 3.4.2,

3.5 Automatic Fake News Detection

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing. The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing.

3.6 Data Pre-Processing

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. Modifies these data individually through pre-processing.

News Processing Steps:

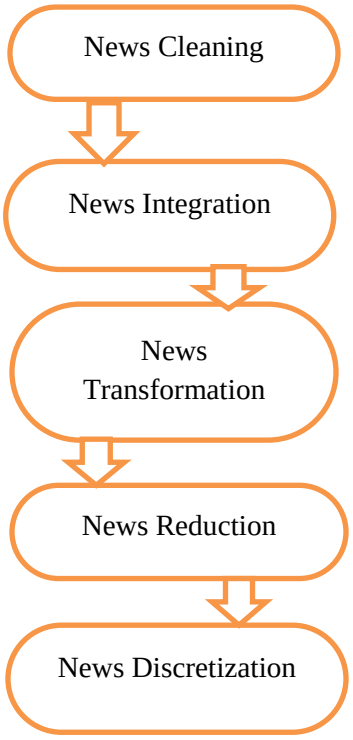


Figure 3.6.1: proposed of News processing

Flow Chart:

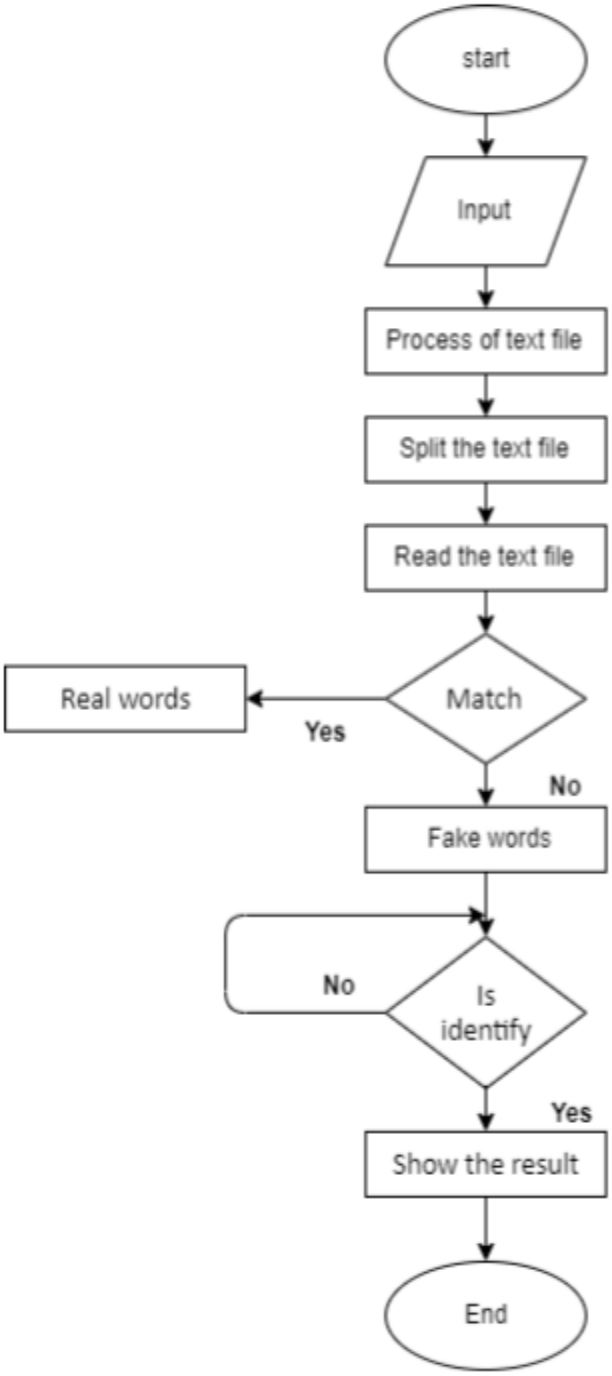


Figure 3.6.2: Proposed flowchart for news identification

3.7 Implementation Requirements

Finally, we have gained a successful result to do this research. And needs some requirements for implementation. Those are bellowed, we ensure that necessary things so. Particularly we achieved using these types of tools are benefited.

Hardware/Software Requirements

- Operating System (Windows 7 or above)
- Hard Disk (minimum 4 GB)
- Ram(more than 1 GB)
- Web Browser(preferably chrome)
- Testing tools

Developing Tools

- Python Environment
- Tensor Flow
- Spyder (Anaconda3)
- Django 1.11 (For UI)
- Notepad++

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This chapter 4 mainly focuses on the descriptive analysis of the data used in the research as well as the experimental results of our project.

4.2 Dataset

4.2.1 Raw Data

Our raw data are from the most renowned news portal CNN, BBC, BuzzFeed, Daily star, PolitiFact, etc., We collect our data by using Corpus. After collecting data, news is stored on text document file. In these file, data are present with some html tag name. So it has become obvious to clean the data. That means pre-processed the row data for preparing for the model.

4.2.2 Cleaning Raw Data

We use a script file to be helpful of our data pre-processing task. This python script file is responsible for:

- Remove all html tag name.
- Remove unnecessary spaces from the text.
- Remove all new line of each news and arrange it in a line.
- Assign an integer number for pre defining the category of each news.

Actually, by this process, we can get all our categorical news in individuals file but the outputted file data are pre-processed and categorical.

4.2.3 Creating Input File

After data cleaning phase, we get six categorical tsv files as we are working on this research on these six categories. The nine categories are: Politics, Crime, Sports, Entertainment, Business, Life Style, Accident, National and International. Hence, after successfully preprocessing process, there have these six categorical news file in our hand. Then, to perform Data Mining Techniques for detecting fake news, we must join all these files into a file. For this, we use another python script named join.py. This file takes the folder name that contains all tsv files as an input and produces only a file where all news contained individually being merged.

4.2.4 Excluded Words Removal

We develop a python code for classify a news into a category. After joining all news into a file, our system is ready for building a model. For this, a little cleaning process is done before. We create a list that contains some Bangla words that actually no related with the category of the news. We called it as Excluded words and named it Excluded words list. Just checking that if excluded words are present in our input file or not. If exists, must be removed.

4.2.5 Feature Selection and Extraction

This phase is the main part of classifying approach and this is feature selection and extraction. It actually, decides, in which perspective classify will be done. We use word count as our feature selection and create it.

4.2.6 Building Model and Fit Dataset for Classifier

To build a model, we separate our dataset into two parts.

- Training Dataset
- Testing Dataset

We use 3:1 ratio for preparing our model. The three portion data set will be treated as training dataset and the rest one portion will be considered as testing dataset.

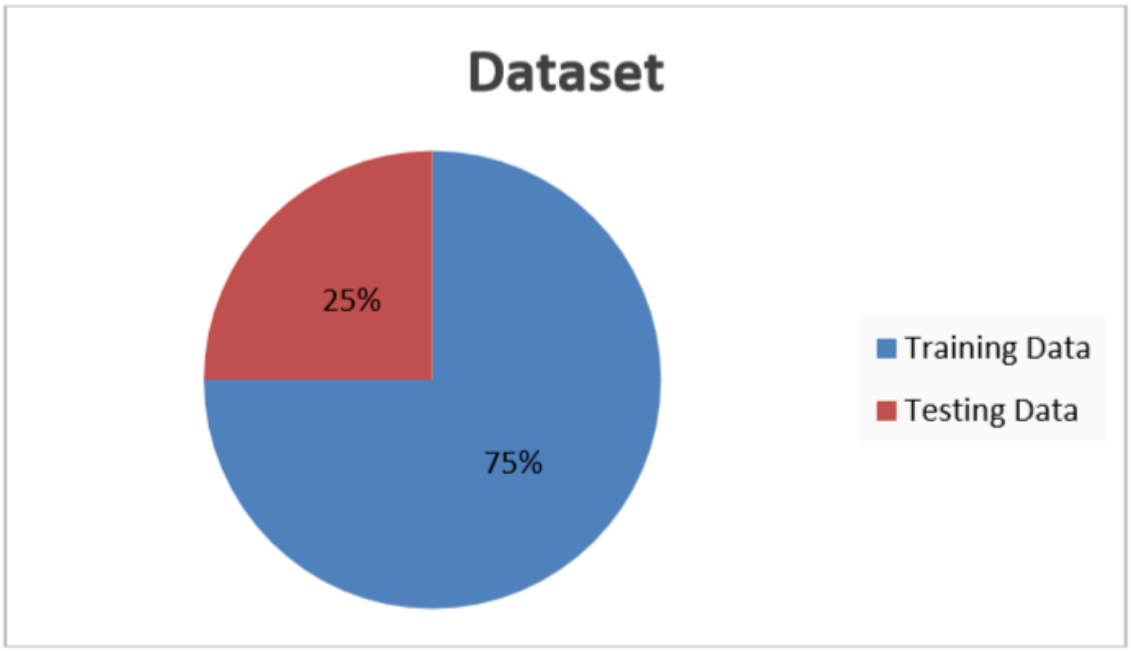


Figure 4.6.1: Dataset chart ratio.

In the concept of percentage, 75% data will be for training and 25% will be for testing. And this will make our expected model, As, we are dealing with several classifiers, we use it by importing sklearn package. This classifier can produce an integer that actually means the category of the expected news.

4.3 False: A New Benchmark Dataset

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are

divided into different numbers. In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources.

Statement	**Speake**	**Context**	**Label**	**Justification**
"The last quarter, it was just announced, our gross domestic product was below zero. Who ever heard of this? It's never below zero."	Donald Trump	presidential announcement speech	Pants on Fire	According to Bureau of Economic Analysis and National Bureau of Economic Research, the growth in the gross domestic product has been belowzero42timesover68years. That's a lot more than "never." We rate his claim Pants on Fire!
"Newly Elected Republican Senators Sign Pledge to Eliminate Food StampProgramin2015."	Facebook posts	Social media posting	Pants on Fire	More than 115,000 social media users passed along a story headlined, "Newly Elected Republican Senators Sign Pledge to Eliminate Food StampProgramin2015." But they failed to do due diligence and were snookered, since the story

				came from a publication that bills itself (quietly) as a "satirical, parody website." We rate the claim Pants on Fire.
"Under the health care law, everybody will have lower rates, better quality care and better access."	Nancy Pelosi	on 'Meet the Press'	False	Even the study that Pelosi's staff cited as the source of that statement suggested that some people would pay more for health insurance. Analysis at the state level found the same thing. The general understanding of the word "everybody" is every person. The predictions don't back that up. We rule this statement False.

Table: 4.3.1: Some random excerpts from the FALSE dataset.

4.4.1 Dataset Statistics

Training set size	**10,269**
Validation set size	**1,284**
Testing set size	**1,283**
Avg. statement length (tokens)	**17.9**

Table:4.4.1 Datasets Statistics

Top-3 Speaker Affiliations:

Democrats	**4,150**
Republicans	**5,687**
None (e.g., FB posts)	**2,185**

Table 4.4.2: The LIAR dataset statistics.

Therefore, it is of crucial significance to introduce a larger dataset to facilitate the development of computational approaches to fake news detection and automatic fact-checking [12]. We show some random snippets from our dataset in Table 4.2.1.

After initial analysis, we found duplicate labels, and merged the full-flop, half-flip, no-flip labels into false, half-true, true labels respectively. We consider six fine-grained labels for the truthfulness ratings: pants-fire, false, barely true, half-true, mostly-true, and true.

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. In this century of digital society, fake news & rumor are the biggest threats because it can easily bring several negative impacts on society. It's very much challenging for readers to differentiate between fake news and real news. Some of the online news portal, blogs & sites who have no proper authorization to publish news but they are continually publishing different types of rumors or worthless news but with spicy headlines to seek the attention of readers which made it challenging to identify reliable and authorized news sources.

4.4 Fake: Benchmark Evaluation

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data.

4.6 Experimental Settings

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data. On each news portal, different types of news are published, together with each of this news we create a Beginning Data Set. Data sets on these data are divided into different numbers. The best filter sizes for the CNN model was (2,3,4). In all cases, each size has 128 filters. The dropout keep probabilities was optimized to 0.8.

Models	Valid	Text
Majority	0.204	0.208
SVMs	0.258	0.255
Logistic Regress0ion	0.257	0.247
Bi-LSTMs	0.223	0.233
CNNs	0.260	0.270

Table 4.6.1: The evaluation results on the LIAR dataset. text-only models.

Hybrid CNNs

Text + Subject	0.263	0.235
Text + Speaker	0.277	0.248
Text + Job	0.270	0.258

Text + State	0.246	0.256
Text + Party	0.259	0.248
Text + Context	0.251	0.243
Text + History	0.246	0.241
Text + All	0.247	0.274

Table 4.6.2: The evaluation results on the LIAR dataset. The bottom: text + meta-data hybrid models.

The first thing we do when we work for research is to pre-process the data. One of the main tasks of data mining is to first modify the data through processing. For this, we collect news from different types of news portals. Then we correct the error by pre-processing the data

.4.7 Results

We outline our empirical results in Table 4.6.2. First, we compare various models using text features only. One of the most important tasks of detecting fake news is to classify the news. Then take these two news stories classified identifiers and analyze their properties. Before that, our biggest motivation is what some news people are looking for more and more, we make those news headlines and perform their work by creating their database. We compare the predictions from the CNN model with SVMs via a two-tailed paired t-test, and CNN was significantly better ($p < .0001$). When considering all meta-data and text, the model achieved the best result on the test data.

CHAPTER 5
SUMMARY

5.1 Summary of the Study

It is not easy to calculate what is wrong and what will be the right news.

Hardly worked that our expected result is so good. We see a lot of research on how can find out more efficient accuracy in their works, and therefore we are applied some efficient algorithms and methods. We calculate the result of how much fake news and how much real news by machine learning approaches. That's the way we got some outstanding results in our research. We mainly focused on social media posted fake news and uploaded content. Ignore confusion to get benefited. Classified the datasets and learned by our own methods that really help us to define the real news and fake news.

5.2 Conclusion

This work, we reached classified text or content in social media, to identify news, which is a fundamental problem for social media mining[9]. We proposed a novel method Trace miner that classifies social media messages and news portals [9]. We described and compared previous datasets and suggest new requirements for future data sets [9]. Making datasets from raw data to accurate data. We compared the accuracy of previous experiments with our related works, we have faced some challenges about the fake news vs. real news identifies.

Very tough to differentiate the types of data sets, and finally we succeeded to get the more efficient accuracy rate of fake news detection.

5.3: Recommendations

That a few recommendations for this are as follows.

- To create the data sets more efficient, to produce a better output of this research work.
- To focus on social media, collect news which was posted in the news portal or social media that can get the proper result.
- Calculate the result using your building algorithms by machine learning approaches.

5.4 Implication for Further Study

One of the most important tasks of detecting fake news is to classify the news. Then take these two news stories classified identifiers and analyze their properties. Before that, our biggest motivation is what some news people are looking for more and more, we make those news headlines and perform their work by creating their database. At last, how to identify low quality or even malicious users spreading fake news is important for fake news intervention.

- Classified the datasets to make this more efficient.
- Used appropriate methods can get good accuracy in this research.
- Defined what types of algorithms make more comfortable to get a good result.

References

[1] Potthast, Martin, Johannes Kiesel, Kevin Reinartz, JanekBevendorff, and Benno Stein. "A stylometric inquiry into hyperpartisan and fake news." arXiv preprint arXiv:1702.05638 (2017).

[2] Tacchini, Eugenio, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. "Some like it hoax: Automated fake news detection in social networks." arXiv preprint arXiv:1704.07506 (2017).

[3] Wang, William Yang. "" liar, liar pants on fire": A new benchmark dataset for fake news detection." arXiv preprint arXiv:1705.00648 (2017)..

[4] Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic detection of fake news." arXiv preprint arXiv:1708.07104 (2017).

[5] Shu, Kai, Suhang Wang, and Huan Liu. "Exploiting tri-relationship for fake news detection." arXiv preprint arXiv:1712.07709 (2017).

[6] Long, Yunfei, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. "Fake news detection through multi-perspective speaker profiles." In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 252-256. 2017.

[7] Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." arXiv preprint arXiv:1811.00770 (2018).

[8] Tschiatschek, Sebastian, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. "Fake news detection in social networks via crowd signals." In Companion Proceedings of the The Web Conference 2018, pp. 517-524. International World Wide Web Conferences Steering Committee, 2018
.

[9] Wu, Liang, and Huan Liu. "Tracing fake-news footprints: Characterizing social media messages by how they propagate." In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 637-645. ACM, 2018.

[10] Ruchansky, Natali, SungyongSeo, and Yan Liu. "Csi: A hybrid deep model for fake news detection." In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 797-806. ACM, 2017.

[11] Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. Journal of Machine Learning Research 12(Aug):2493–2537.

[12] Koby Crammer and Yoram Singer. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research 2(Dec):265–292.

[13] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. Association for Computational Linguistics, pages 171–175.

[14] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL.

[15] Alex Graves and J¨urgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks 18(5):602–610.

[16] Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, XiaoLi Li, Guangxia Li, and Ant Financial. 2016. Deceptive review spam detection via exploiting task relatedness and unlabeled data. In EMNLP.

[17] Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. Neural computation 9(8):1735–1780.

[18] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pages 309–319.

[19] Ver´onica P´erez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In EMNLP. pages 1120–1125.