Shahjalal University of Science and Technology
Software Engineering
Institute of Information and Communication Technology

SWE 450

BanFakeVis: A Textual and Observational Study of Bangla Fake News in the Era of Covid-19

Ahsan Aziz Ishan
Reg. No.: 2016831017
4^{th} year, 2^{nd} Semester

Kazi Tushita Tahsin
Reg. No.: 2016831026
4^{th} year, 2^{nd} Semester

Software Engineering
IICT, SUST

Supervisor
Asif Mohammed Samir
Assistant Professor
Software Engineering
IICT, SUST

December 25, 2023

BanFakeVis: A Textual and Observational Study of Bangla Fake News in the Era of Covid-19

A Thesis submitted to the Department of Software Engineering,
Shahjalal University of Science and Technology, in partial fulfillment of the requirements for the degree of B.Sc.(Engg.) in Software Engineering.

Ahsan Aziz Ishan
Reg. No.: 2016831017
4^{th} year, 2^{nd} Semester

Kazi Tushita Tahsin
Reg. No.: 2016831026
4^{th} year, 2^{nd} Semester

Software Engineering
IICT, SUST

Supervisor
Asif Mohammed Samir
Assistant Professor
Software Engineering
IICT, SUST

December 25, 2023

Recommendation of the Thesis Supervisor

The thesis entitled "BanFakeVis: A Textual and Observational Study of Bangla Fake News in the Era of Covid-19" submitted by the students

1. Ahsan Aziz Ishan
2. Kazi Tushita Tahsin

is under my supervision. I, hereby, agree that the project can be submitted for examination.

Signature of the Supervisor:
Name of the Supervisor: Asif Mohammed Samir
Assistant Professor
Software Engineering
Institute of Information and Communication Technology
Shahjalal University of Science and Technology, Sylhet
Date: December 25, 2023

Certificate of Acceptance of the Thesis

The thesis entitled "BanFakeVis: A Textual and Observational Study of Bangla Fake News in the Era of Covid-19" submitted by the students

1. Ahsan Aziz Ishan
2. Kazi Tushita Tahsin

on December 25, 2023 is, hereby, accepted as the partial fulfillment of the requirements for the award of their Bachelor Degrees.

Director of IICT	Chairman, Exam. Committee	Supervisor
Prof Dr. M. Jahirul Islam	Prof Dr. M. Jahirul Islam	Asif Mohammed Samir
PhD, PEng	PhD, PEng	Assistant Professor

Abstract

The Covid-19 pandemic is one of the largest pandemics to hit the 21st century affecting not only the lives of the people who are infected, but also the people living in lockdown as the digital screen time for people has increased significantly in a short amount of time. The greatest challenge the online community has faced during this crisis period has been dealing with the circulation of misinformation online which has even been the cause of death in worst cases. The Bangla language is no exception to this. For this reason, we attempted text analysis and observational study which include News Categorization, Word Count, Parts of Speech tagging(POS), Named Entity Recognition(NER), Sentiment Analysis, Punctuation and Image Count in our collected data set of Bangla fake news articles. Our study succeeded in achieving some satisfying results in Word, Punctuation and Image Count that we hope would be of some help in the field of Bangla fake news and text analysis in the future.

Keywords **:** Covid-19,Bangla fake news, Fake news, News Categorization, Word Count, Parts of Speech tagging(POS), Named Entity Recognition(NER), Sentiment Analysis

Acknowledgements

We would like to thank the faculty of Software Engineering,IICT, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh, for supporting this research. We are very thankful to our honorable supervisor Asif Mohammed Samir, Assistant Professor, SWE, IICT whose contributions have been invaluable for our research.

At last, we are thankful to our thesis committee because all our academic achievements depend on their sacrifice.

Dedication

We would like to dedicate our research to all the online fact checkers and front line workers of Covid-19.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Fake news , if people believe, related to COVID-19 increases anxiety, stress, and even depression. The reason behind this is to stay-at-home, curfews, and closing of essential businesses. Many families are now phasing unemployment scenario. Some are forced to work from home due to the direction of companies. Many companies and businesses are working with loss due to coronavirus. This new life is stressful enough. So anxiety of people is clearly increased due to new life style.

Fake news is primarily spread via social media platforms, such as Facebook, Twitter, etc. Fake news is hard for some people to identify and can create confusion about true fact. People has doubt about accurate information. For example, When consumers learn information via fake news that the particular product is long-lasting. They have curiosity to see and purchase the product. Since it is not possible to purchase during lockdown or the information is incorrect. It made distress on consumers.

Researches have been going on to study the way fake news can be identified and detected. It is true that the main way to combat fake news is by educating ourselves to understand fake news. Machine learning method may help to get identify false information during COVID-19. Facebook and Twitter are now taking measures to remove fake news and misinformation from their platforms. The main objective of the study is to make sure that the information received from social media/mobile/internet are reliable and true. COVID-19 pandemic in 2020 dominates the media, both domestically and abroad. Alongside it is required for the attention on the pandemic first as well as to stop viral spread of fake news online related to coronavirus. At the time of COVID-19, the spread of fake news offers unique challenges and dangers to the public. This paper analyzes 90 fake news and real news data-set related to COVID-19 to

visualize the characteristics of both of them. Also we conducted a survey in the social media platforms to find how people see fake news and how much they trust the news publishers after being victim of disinformation in the era of COVID-19.

1.1 Motivation

COVID-19 affected us all in many ways. The term 'INFODEMIC' describes how scarcity of information made this pandemic situation even worse. A lot of people believed in disinformation that spread during COVID-19 and helped in spreading those information even more. Bangla news web portals, and social media sites played a vital role spreading those disinformation. There have a little research on COVID related Bangla fake news. We tried to collect as much as fake news we could collect from fact-checking websites and social media. So a proper analysis on this fake news is a need of time that has driven us to research on
"BanFakeVis: A Textual and Observational Study of Bangla Fake News in the Era of Covid-19"

1.2 Report Structure

- Chapter 2: Literature Review, also described as background study. Contains short description of all the related works for this paper.
- Chapter 3: The details of data-sets with necessary statistics are mentioned in full details in data analysis section
- Chapter 4: Methodologies of the Data Analysis and Survey
- Chapter 5: Results and Findings of the Analysis and Survey.
- Chapter 6: Conclusion and Future Work.

Chapter 2

Literature Review

Researchers have been searching for ways to cope with fake news circulation online for some time now. In recent times, Twitter and Whatsapp has even updated their policy for coping with malicious news content. There are plenty of literature that led to this report on Covid-19 related fake news analysis.

2.1 Fake News Detection

Neural Networks have been used the most in previous works of fake news detection. One work suggests using an Event Adversarial Neural Network(EANN) for fake news detection that uses Convolutional Neural Network(CNN) and Visual Geometry Group 19 (VGG 19) to detect fake news using Twitter and Weibo data sets [2].

Geometric deep learning has also been used to detect fake news and has an accuracy of 92.7% [3]. The data used here has been collected from fake news debunking sites and this model uses 4 layers of CNN to detect fake news. This model generates graphs using the cascade the news follows and URL of the news and ages well

Another approach that has been used to detect fake news in the past is the Naive Bayes Classifier which classifies the probability of a news type by taking into account some common properties of new articles [5]. There are also studies showing the difference in behavior between Twitter bots and human users during Covid-19 pandemic that can be of help in detecting fake users hence fake news [6].

2.2 Sentence Interaction and Structure Analysis

It has been proven in previous research that that different news types have different sentence structures.But the people of the School of Computer Science of Carnegie Mellon University have used Convolutional Neural Network(CNN), Long Short Term Memory(LSTM) and BERT to make a Graph Attention Network(GAT) which generates an attention map showing the sentence interactions in different types of fake news, making a point indicating that sentences also interact differently in different news types. [27].

Regarding sentence structure, Hierarchical Discourse Level Structure have been used to detect and analyse fake news [28]. This model shows relationship between defferent sentence features like Parent-child distance, number of leaf nodes and points out that the results for fake news are vary a lot from real news. The base of this type of analysis is the Rhetorical Structure Theory(RST) that is a theory of text organization describing the relations that hold between the parts of a text. To develop a RST treebank, the sentences of a data set have to be annotated. There are some annotation guidelines about how to annotate Bangla sentences [30] to develop such treebanks to guide researchers in developing a Bangla RST Treeabank [29].

2.3 Theories Regarding the Spread of Misinformation

Various theories are existing regarding the motivation, reason and perspective behind propagation of misinformation during Covid-19. One paper suggests that the technique used for spreading fake news is similar to a theory given by philosopher Louis Althusser which states that, "The state uses verious kinds of ideological mediums to manipulate the common people to obey the state." This is known as the Ideological State Apparatus [19].Furthermore, another paper proposes that the psychology of the people spreading these type of news voluntarily involves the belief that, "Whoever controls the media controls the mind" [20]. This is a very useful belief to have in a time where peoples digital screen time has increased dramatically compared to before. This type of news spreading often leads to social entropy leading to stigma and violence. One of the papers give some reasoning as to why people share misinformation so often nowadays. From their perspective, this happens because people tend to online news portals over traditional ones too much now, because of information overload and Cyberchondria [10]. Younger people are more likely to share misinformation than older people because of their digital screen time ex-

posure difference. A similar paper suggests that for this Indian subcontinent, fake news spreads rapidly because the people here are media illiterate, there is too much information to grab and the communication in the 21st century in social has become 2 way [16].

The Bruno Kessler Foundation analysed 112 million public social media posts and came to a conclusion that the motivation behind the deliberate spread of Covid-19 related misinformation is speculated to be pure mischief, politics or economics [21].

An anlysis on 150 sns users content in a study show that Covid-19 related misinformation spread the most in March via Twitter, in April via email, in May via email and Twitter and in June via email and mobile phones [15]. Some are other studies also talk about the propagation of misinformation during the Covid-19 pandemic and natural disasters [3] [17].

Some policy frameworks are proposed by studies which include tranperancy of news, best possible treatment in online, availabulity of latest technology for frontline workers, verifying news before sharing, developing government platforms to fight against this misinformation propagation etc [11] [17] [19].

2.4 Fake News Categorization

News categorization is also a topic for analysing news. There are studies giving an exact insight on which topics should be a part of disinformation and which topics should not [23]. A study done on 20.8M Covid-19 related fake tweets categorizes their data set into unreliable, conspiracy theories, clickbait and political or biased categories [7]. A similar study done using data from several social media further classifies Covid-19 related fake news into rumor, stigma and conspiracy theory where 89% found out to be rumor and also in true, false, misleading and not proven categories where 82% data is proven to be false [8].

There is also a study done with only Covid-19 related fake Youtube content including Covid-19 related information, news updates and personal experiences where it was found that 10% of the content was misleading [14].

A study from the U.S consulate in Kolkata and South Asian Forum for Environment in Kolkata categorizes misinformation spread during environmental disasters as misinformation, disinformation and malinformation [17]. Tri-grams are said to be helpful in classifying Bangla text into categories [9].

2.5 Analysing Fake News on Different Features

There has been plenty of work done regarding analysis of fake news content and title that are noteworthy. Covid-19 related Twitter conversations have been analysed based on their sources, following trends and their sentiment and topic [7]. Another analysis in Facebook in Italy during the Covid-19 period shows the trend in the controversial topics regarding Covid-19 and the propagation pattern of the news URL.

There have also been studies using deductive coding on social media contents regarding the Covid-19 misinformation that analyses the theme, media type, source, international coverage and intention of the contents found [18]. It goes on to show that majority of the fake news articles are text and video combined.

A survey paper concludes that most fake news can be detected by analysing their knowledge type, writing style, propagation patterns and their source credibility [1]. Some studies take into account the relationship among news publishers, news content and news readers to be key factors for the propagation of fake news [4]. Another study analyses Covid-19 related fake news from different fact checker sites based on their format, source and claims [13].

One study proposes some interesting analyzing features that include the use of acronyms and initialisms, word reduction, phonetic letters and numbers, stylized or unconventional spelling, emoticons, stylized or unconventional punctuation and a large number of images to be features to be analyzed for the detection of fake news [26].

2.6 Analysing Fake News using Various Analysing Techniques

Some studies used Word Count and Parts of Speech tagging to show the different patterns in Covid-19 related fake and real news [22]. In End to End Parts of Speech Tagging and Named

Entity Recognition in Bangla Language, researchers used a BLSTM-CNN-CRF based model for POS tagging and NER that achieved an accuracy of 93.86% in POS tagging [24].

Fake news can also be seen to have a lot going on in the headlines and uses Superlatives, Comparatives, Adverbs, Personal Pronouns and Hedging Words frequently [25].

Chapter 3

Data Collection and Pre-processing

A dedicated Bangla news dataset involving only fake Covid-19 news could not be found. Hence we tried to create a dataset of our own consisting of 90 Bangla fake news and 90 Bangla real news from various sources. A survey was also conducted to find out the public perspective regarding fake news dduring this crisis.

3.1 Data Source

3.1.1 News Data

The fake news dataset has been collected from 46 different news sites, Facebook and Youtube with the help of some fact-checking websites.While the real news dataset has been collected from 13 different news sites. We collected data from both renowned news sites such as Prothom Alo, Bdnews24 as well as from less known sites from October 2020 till January 2021. All data has been collected manually and verified. After collection, data has been converted into csv files to use in algorithms. No satire news has been used in making this data sets.

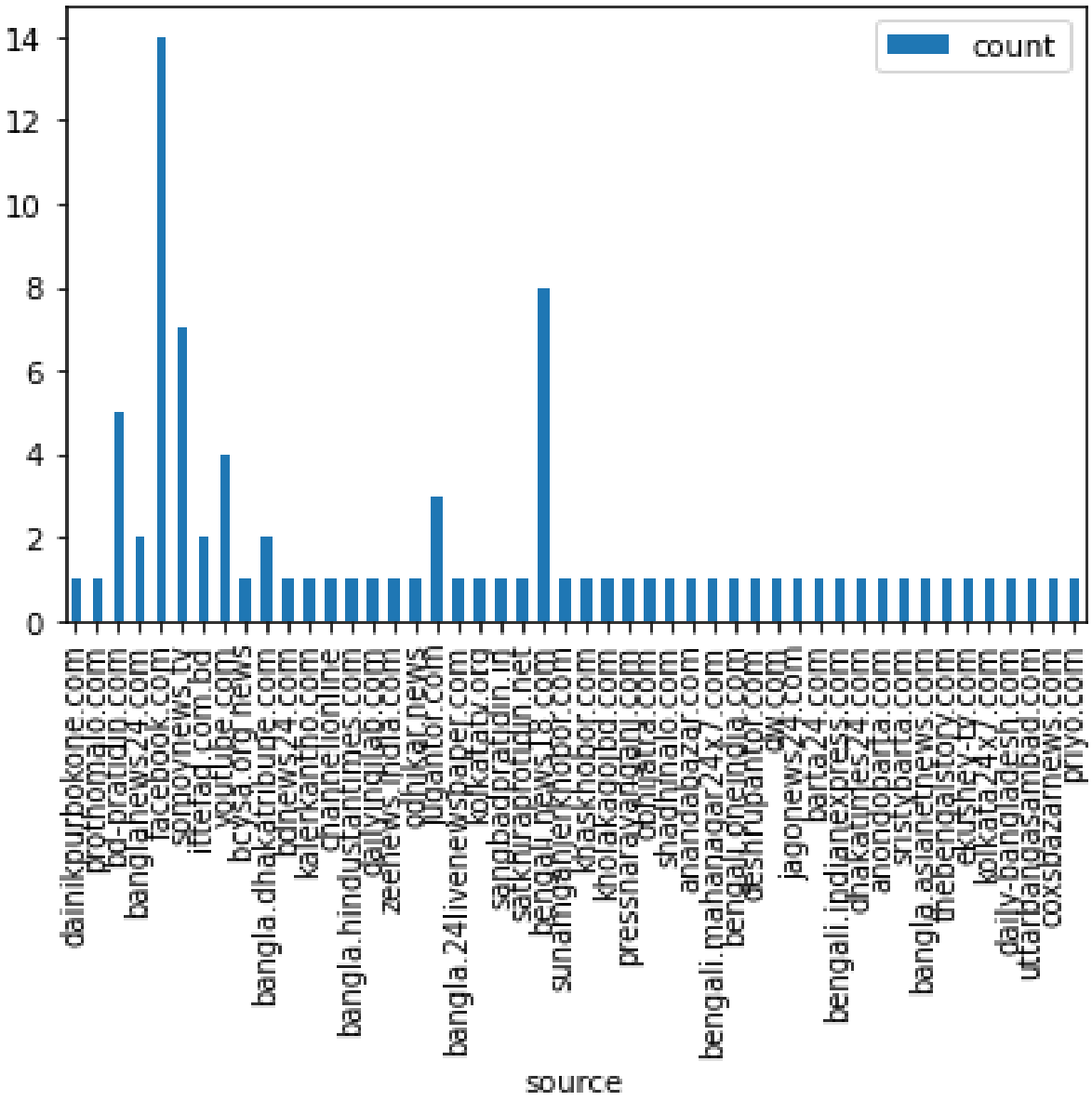

Figure 3.1: Fake News Sources

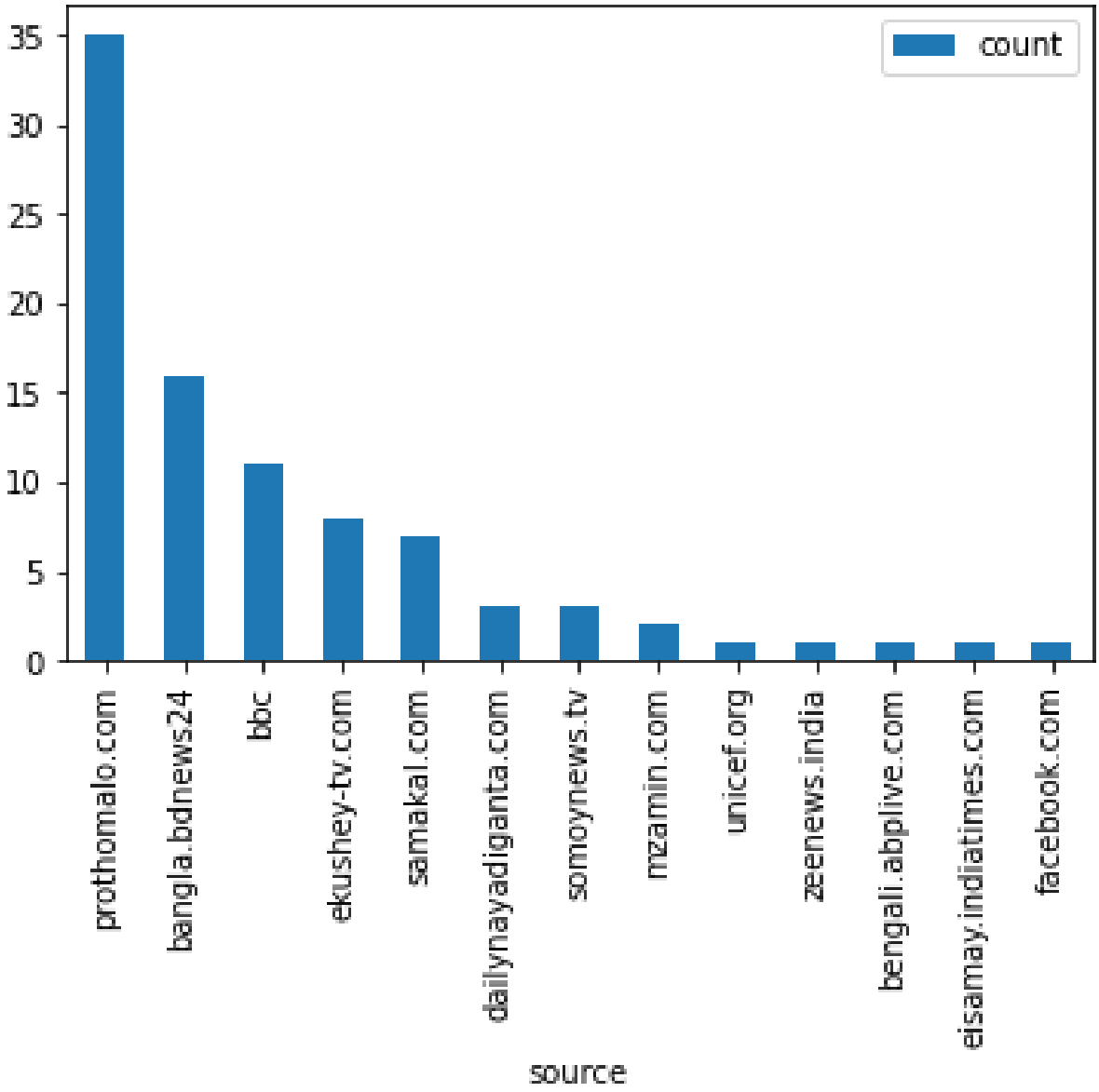

Figure 3.2: Real News Sources

3.1.2 Survey Data

The survey data has been collected by conducting an online survey at Facebook where we got a total of 152 response for 5 questions that were set as a google form. This survey was conducted in December 2020.

1. If you see an online news, do you read only the headlines or whole news ?

2. Among these which news portals do you think are trustable ?

3. Have you ever seen sharing clickbaits or fake news in the news portals you've selected ?

4. Have you ever shared a fake news (By mistake, or not verifying truly from social media that was later proven to be fake) ?

5. Do you verify news before sharing ?

3.1.3 Sentiment Analysis Data

The data-set used in sentiment analysis is created at Shahjalal University of Science and Technology [31]. This data-set contains of 18903 sentences and the sentiment in them.

3.2 Data Pre-processing

The news data needed to be processed as without processing, it was not possible to use the data for the various analysing techniques.

The data cleaning techniques that have been used in this work are -

- Removing stop words:
  Stopwords are the words in any language which does not add much meaning to a sentence. Bangla stop words were removed so that we could the most meaningful frequent words when we counted word frequency for the news articles.

- Stemming:
  In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. This was also done for counting the word frequency and tf-idf.

- Removing punctuations:
  Removing punctuation was also important for counting the word frequency of the articles and for tf-idf.

- Removing Line Breaks:
  Since news articles generally include several paragraphs, all line breaks in all articles were removed for analyzing data smoothly.

Chapter 4

Methodology

The analysis on fake news regarding Covid-19 has been done using two datasets. One is the dataset we collected from various websites and the other one is one that was collected through an online survey. Hence our methodology for the analysis is divided into two parts: Analysis on fake Bangla Covid-19 news reports and Analysis on Online survey.

4.1 Analysis on Fake Bangla Covid-19 News Reports

In this part, we applied some language analysing techniques on the data sets that we created manually consisting of 90 Bangla fake news and 90 Bangla real news. The purpose for conducting this was to unveil what language features played important roles in making a news fake, to find if there is any dependency among the features and comparing the results we get for the fake news data set with the real news data set to spot the differences among them.

4.1.1 News Categorization

Categorization is the action of classifying things into a categories. News can be of many categories. For example: National news, sports news etc. various studies suggest various categories according to which our data set could be classified [7] [8] [14] [17] [18] [21]. Regarding Covid-19, the fake news that we collected could be categorized based on what type of information they are and what field they relate Covid-19 with.

4.1.1.1 News Categorization based on Information type

Fake news is false or misleading information presented as news. Fake news can be of various types. Some of them are -

- False News:
  False news is news that contains information that is entirely false.

- Controversial News:
  Controversial news are those that contain a topic over which the public opinion is conflicted.

- Clickbait:
  On the internet, clickbait news is content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page, it may or may not the contain the news that is written in the title

- Misleading or Deceptive News:
  Misleading news are news that are written to mislead people and it is done deliberately most of the time.

- Fabricated News:
  Fabricated news are news articles whose entire story is fabricated by someone and made to look like a real news article by including some subtle truths in places.

- Rumors:
  A rumor is a a currently circulating story or report of uncertain or doubtful truth.

- Disinformation:
  Disinformation is false information which is intended to mislead, especially propaganda

issued by a government organization to a rival power or the media.

- Cherry Picking:
  In news context, cherry picking is the action or practice of choosing and taking only the most beneficial parts of a news from what is available. It refers to not publishing the entire truth about something but only publishing the parts by which the agency is going to get more clicks and benefit.

  The fake news data set was first categorized taking these fake news categories into account and the results were plotted in a bar graph. There are 9 categories in total.

4.1.1.2 News Categorization based on Content

The Covid-19 related fake and real news content that we found regarding Covid-19 includes different topics that represent Covid-19 from different perspectives. Based on the news content, both the fake and real news data were categorized and potted in bar graphs to compare which type of news content ruled Covid-19 related fake and real news.

4.1.2 Word and Sentence Count: Total and Average

The total number of words are counted for each of the data set to figure out if there is a difference in length in Covid-19 related fake news and real news data. This word count is done after some data cleaning and processing which includes -

- Removing stop words from the data sets
- Removing punctuation
- Removing line breaks
- Stemming
- Word tokenization which is the process of splitting a large text sample into words so that we can count each word invidually

One of the papers suggested that fake news has shorter content length than real news but packs a lot in tis title [26]. To test theory is practice -

- We counted the total and average number of words in both the fake and real new data sets after tokenization. The words were counted separately for news titles and content
- The same thing was done with sentences after segmenting the titles and contents of the data sets separately

4.1.3 Parts of Speech and Adjectives,Adverbs, Pronouns and Hedging Words Tagging

4.1.3.1 POS Tagging

Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context. The goal for POS tagging in our data sets were to find out trends of parts of speech use in Covid-19 related fake and real news.POS tagging has been used in some studies before to analyse fake news [22] [24] [25]. A POS tagger generally has a trained model that has been trained with supervised tagged data. It tokenizes the given text data into words, then fits the data into its pretrained model and tags them. It's one kind of prediction. We used the bnlp POS tagger for tagging our data sets. Some of the acronyms used in POS tagging are CC meaning Coordinating Conjunction, DT meaning Determiner, CN meaning Common Noun etc.

4.1.3.2 Adjectives, Adverbs, Hedging Words and Pronoun Tagging

Since the results of POS tagging were not satisfactory, we decided to go with an alternate approach which included searching for the frequency of specific parts of speech that are known to be prominent in fake news articles [22]. This includes Superlatives(JJS), Comparatives(JJR), Hedging Words(HW), Personal Pronouns(PRP) and Adverbs(RB).

Comparative Adjectives (JJR)	চেয়ে, চাইতে, হতে, অপেক্ষা, থেকে
Superlative Adjectives (JJS)	সবচাইতে, সবথেকে, সর্বাপেক্ষা, সর্বাধিক, সবচেয়ে
Personal pronoun (PRP)	নিজেই, আমরা, নিজেকে, নিজে, আমি, নিজেদের, নিজেরা, আমার, আমাদের তোমার, তোর, তোমাদের, তোদের, তুমি, তুই, তোমরা, তাকে, তার, তাদের, তাহাদের, তাহার, আপনি, আপনার, আপনাদের, আপনিই, আপনাদেরই, তাদেরই, নিজেরই, নিজেদেরই, নিজেকেই, তোদেরই, তাহাদেরই, তাহারই, তোমাদেরই, তোরই, তোমারই
Hedging words	মনে হয়, হতে পারে, সম্ভবত
Adverbs (RB)	বেশি, অনেক, খুব, খুবই, অতিশয়, অতি, একটুও, পুরও, সম্পুর্ন, পুরোপুরি, সব, সত্যি, প্রকৃতই, তারাতারি, এত

Figure 4.1: Adjectives, Adverb and Hedging words

Since we could not find a definite list of Bangla words of these parts of speech, we made one ourselves using the resources we could find online.

The data sets were first tokenized, then we iterated over the data to count the number of adjectives, adverbs, personal pronouns and hedging words in the data. We used for loop to loop through the data as the list of words that we managed to collected was very small. This was done for both Covid-19 related fake news and real news data for comparison.

4.1.4 Named Entity Recognition (NER)

Named entity recognition (NER) , also known as entity chunking/extraction , is a popular technique used in information extraction to identify and segment the named entities and classify or categorize them under various predefined classes.An NER System is capable of discovering

entity elements from raw data and determines the category the element belongs to.It helps to easily identify the key elements in a text, like names of people, places, brands, monetary values, and more. This is also a technique that is used in fake news analysis [24].

An NER tagger and a POS tagger works in a similar way. It also tokenizes the given text data,fits the data into a pretrained model that has been trained with supervised meaning tagged NER data and then tags the data. The NER tool we used here is the NER tool fron bnlp toolkit which recognizes Person, Location, Object and Organization in the beginning, inside, end and as a single item. Again, we used NER to observe what kind of words were used in Covid-19 related fake and real news and to find out the relationship between them.

4.1.5 Unconventional Use of Punctuation

Punctuation marks are used in a sentence to give it structure, for the ease of reading, to convey different meanings of a sentence and to express emotions. There are mainly 14 kind of punctuation marks in a language. We tried to identify patterns in which some certain punctuation marks are used in news articles to make it stand out from rest of the news articles. The motivation of this was found in a paper that has proven this to be true for English news [26].

- We took '!', '?' and '.' in an array and iterated through the contents of fake news and real first
- After that, we took the same array and iterated only through the titles of the data sets

The results are similar for both the titles and contents of the data sets.

4.1.6 Number of Images in a News Article

A picture is a great way to deliver a message in an instant as it can be remembered and recalled easily. Images have the power to move people emotionally as human brain processes image quickly. According to studies, nowadays mainstream media is all about manipulating people emotionally for their own benefit. So, a news with less facts but with an emotional picture gets more attention than a news with only facts as our brain processes the image first

and is immediately clickbaited [26].

On the basis of this theory, we decided to count the number of images per news article in our data sets to test whether Covid-19 related fake news also follows this pattern or not.

4.1.7 Sentiment Analysis on News Headlines

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is done for analysing news articles to find out their motivation behind spread [6] [7].Sentiment analysis algorithms fall into one of three buckets:

- Rule-based: these systems automatically perform sentiment analysis based on a set of manually crafted rules.
- Automatic: systems rely on machine learning techniques to learn from data.
- Hybrid systems combine both rule-based and automatic approaches.

There are two stages involved in implementing automatic systems:

- Training
- Prediction

In the training stage, a sentiment analysis model learns to correctly tag a text as negative, neutral or positive using sample data. The feature extractor then transforms the text into a feature vector, creating pairs of feature vectors and tags (e.g. positive, negative, or neutral) that are fed into the machine learning algorithm to generate a model. In the prediction process, the feature extractor is used to transform unseen text into feature vectors, which are fed to the model, enabling it to make sentiment predictions.

We used LSTM based RNN for our sentiment analysis. The idea behind RNNs is to make use of sequential information. In a traditional neural network we assume that all inputs (and outputs) are independent of each other. If we want to predict the next word in a sentence we better know which words came before it. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous

computations and we already know that they have a “memory” which captures information about what has been calculated so far.
LSTM is a special kind of RNN’s, capable of Learning Long-term dependencies. LSTM’s have a Nature of Remembering information for a long periods of time is their Default behaviour. RNN are not able memorize data for long time and begins to forget its previous inputs. To overcome this problem of vanishing and exploding gradient LSTM is used. They are used as solution for short term memory learning. Also in RNN when a new information is added RNN completely modifies the existing information. RNN is not able to distinguish between important or not so important information. Whereas in LSTM there is small modification in existing information when a new information is added because LSTM contains gate which determine the flow of information.

We used a dataset created at Shahjalal University of Science Technology to train our model for sentimental analysis [31]. The dataset contains 13803 sentences. We splitted our dataset into 0.8-0.2 ratio for training and testing.

To train the model, we tokenized the dataset and converted text to sequences. Then we padded our sequence so that all text vector have same length. We defined our input dimensions as 5000 and output 256. We applied the dropout 0.3 to the input, The Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting. Inputs not set to 0 are scaled up by 1/(1 - rate) such that the sum over all inputs is unchanged. Activation function softmax was used to train the model. Finally we had batch size of 32 and epoch size of 8 to train our model with 84 percent accuracy.

Using the model we trained, we analysed the sentiments of our fake and real news. We preprocessed our collected news, tokenized the sentences to analyze the sentiment in sentence level.

4.2 Analysis on Online Survey

A survey is a research method used for collecting data from a predefined group of respondents to gain information and insights into various topics of interest. They can help gauge the representativeness of individual views and experiences. When done well, surveys provide hard numbers on people’s opinions and behaviors that can be used to make important decisions so surveys have been done in previous studies [10] [16]. We conducted an online survey on Face-

book to get peoples perspective on some questions regarding the reliability of the news sites they trust. We got 152 responses. The 5 questions that we set in the survey is shown in Fig - 4.2

1. If you see an online news, do you read only the headlines or whole news?	a. Only Headlines b. I click the news to read full news c. Sometimes I read the full news
2. Among these which news portals do you think are trustable?	☐ Bdnews24 ☐ প্রথম আলো ☐ Somoy TV ☐ দৈনিক পূর্বকোণ ☐ কালের কন্ঠ ☐ যুগান্তর ☐ CNN Bangla ☐ Bangla Tribune ☐ Ananda Bazar ☐ Priyo.com ☐ None of these
3. Have you ever seen sharing clickbaits or fake news in the news portals you've selected?	a. Yes b. No c. Maybe
4. Have you ever shared a fake news (By mistake, or not verifying truly from social media that later proven to be fake)?	a. Yes b. No c. Maybe
5. Do you verify news before sharing?	a. Yes b. No c. Maybe

Figure 4.2: Questions used in the survey

We've collected some data from the participant group including their age range and sex shown in 4.1:

Total Response	152	
Male	102	67.333%
Female	36	23.3333%
AGE		
Range	Boys	Girls
0 - 17	1 (0.67%)	0
18-25	95 (63.33%)	32 (21.33%)
25-35	4 (2.67%)	0
36-45	0	1 (0.67%)
46-above	1 (0.67%)	2 (1.33%)

Table 4.1: Age and Gender data of Participants

Chapter 5

Results and Findings

5.1 Results

5.1.1 News Categorization Based on Information Type

Among the 90 fake news regarding Covid-19 in our dataset, 27% are false, 26% are either false or rumor, 16% are disinformation, 15% are controversial, 9% are misleading and the other 7% include clickbaits and other types of news.

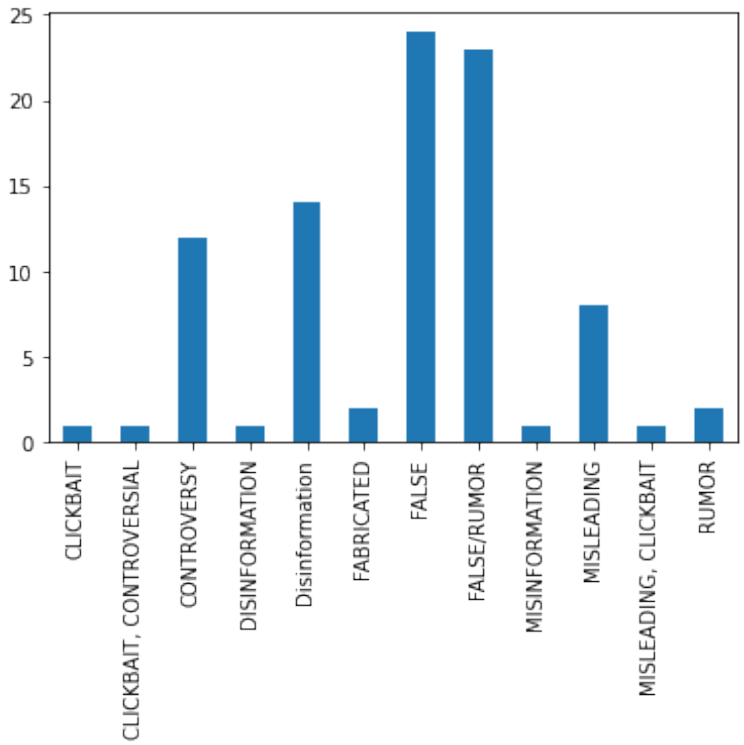

Figure 5.1: Category of fake news (1)

5.1.2 News Categorization based on Content

For Covid-19 related fake news, the results show that 22% are medicine related, 14% news are prophecies, 13% are food related pointing out which foods are good for preventing Covid-19, 9% are vaccine related and both pet and animal related and politics related news are 8%.
In Covid-19 related real news, 19% of the data is vaccine related, both social entropy and knowledge related data are 13%, both Covid-19 related advice and number of Covid-19 case related data are 11%, 9% data is political and 8% data is Lockdown related.

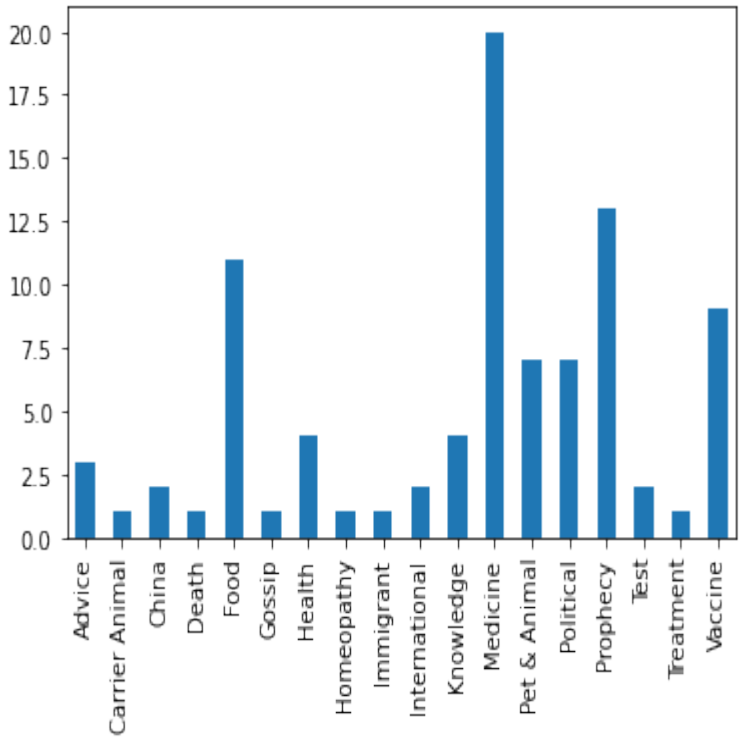

Figure 5.2: Category of fake news (2)

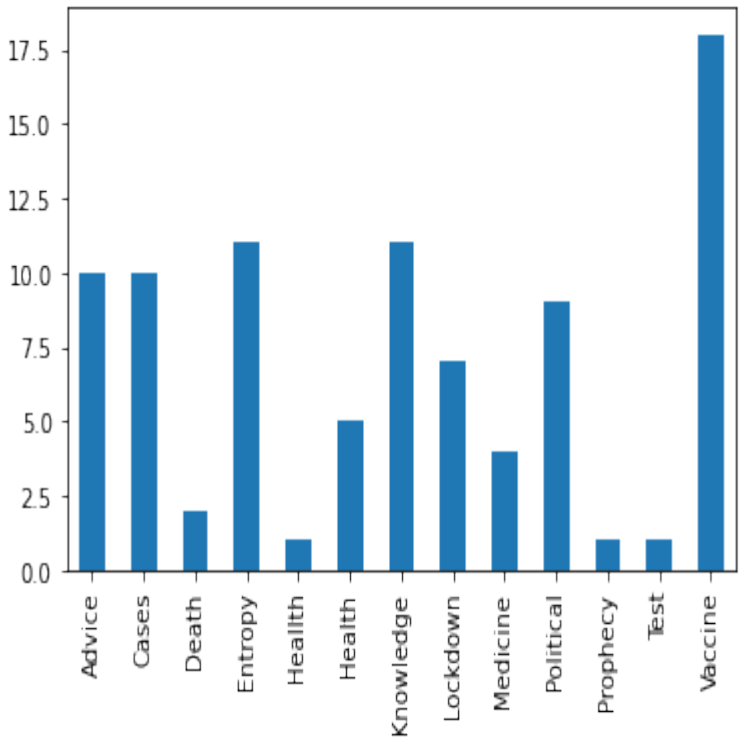

Figure 5.3: Category of real news

5.1.3 Word and Sentence Count: Total and Average

The total number of words in Covid-19 ralated fake news headlines is 871 which in average is 10 words per headline and for real news headlines it is 666 words in total which in average is 7 words per headline.

For fake news content, the number of total words is 30429 which is 338 words in average per news and for real news there are 36422 words in total whose average is 405 words per news.

The total number of sentences in fake news headlines is 100 while in real news it is 95. Both of them have the same average of 1 sentence per headline.

News Headline Analysis		
	Fake	Real
Total Sentence Tokens	100	95
Average Sentence Tokens	1	1
Total Word Tokens	871	666
Average Word Tokens	10	7

Table 5.1: News Headline Analysis

But, the total number of sentences in fake news content is 1863 which is 20 sentences in average per news while in real news content, the total number is 2309 which has an average of 25 sentences per news.

News Content Analysis		
	Fake	Real
Total Sentence Tokens	1863	2309
Average Sentence Tokens	20	25
Total Word Tokens	30429	36422
Average Word Tokens	338	405

Table 5.2: News Content Analysis

5.1.4 Parts of Speech Tagging

In Covid-19 related fake news, 41.4% is Common Noun(NC), 10.9% is is Finite Verb(VM), 6.9% is adjective(JJ) and 5.7% is Proper Noun(NP).

Simlarly in real news data, 42.7% is Common Noun(NC), 10.2% is Finite Verb(VM), 7.3% is Adjective(JJ) and 5.9% is Proper Noun(NP).

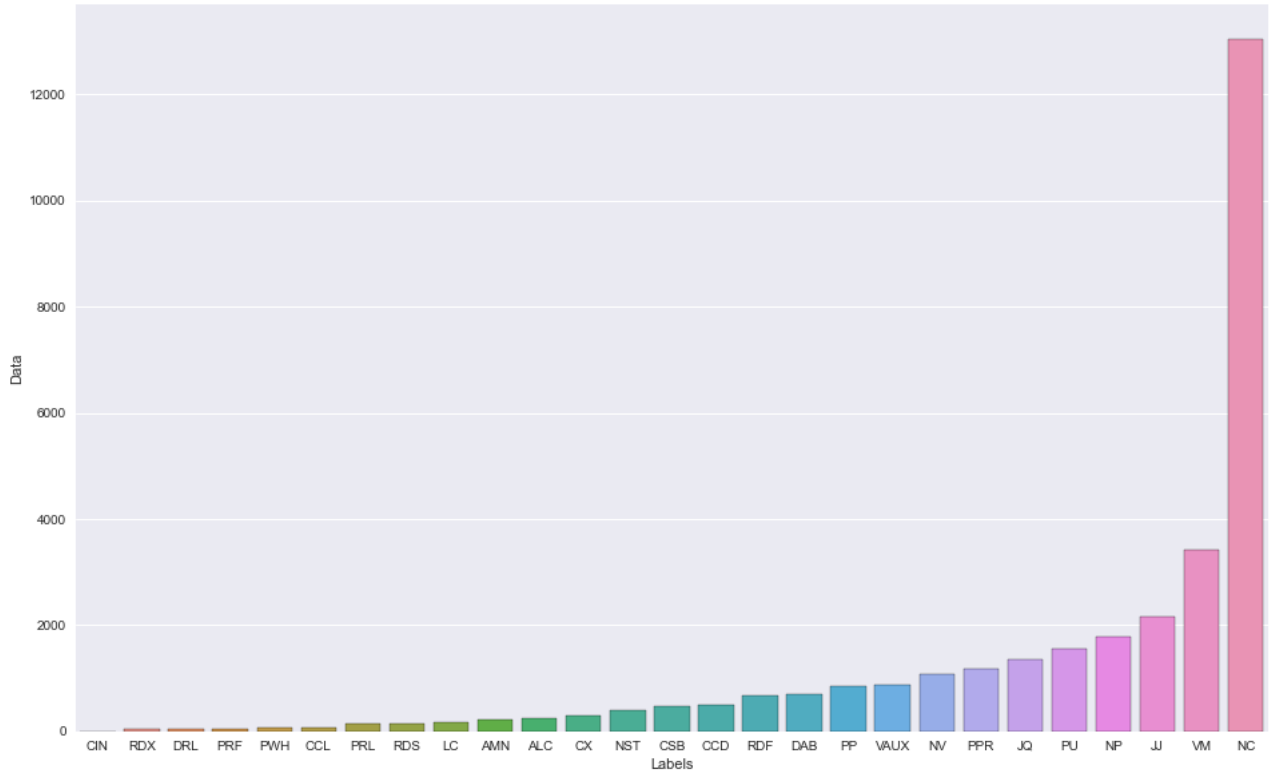

Figure 5.4: Fake News Parts of Speech Tagging

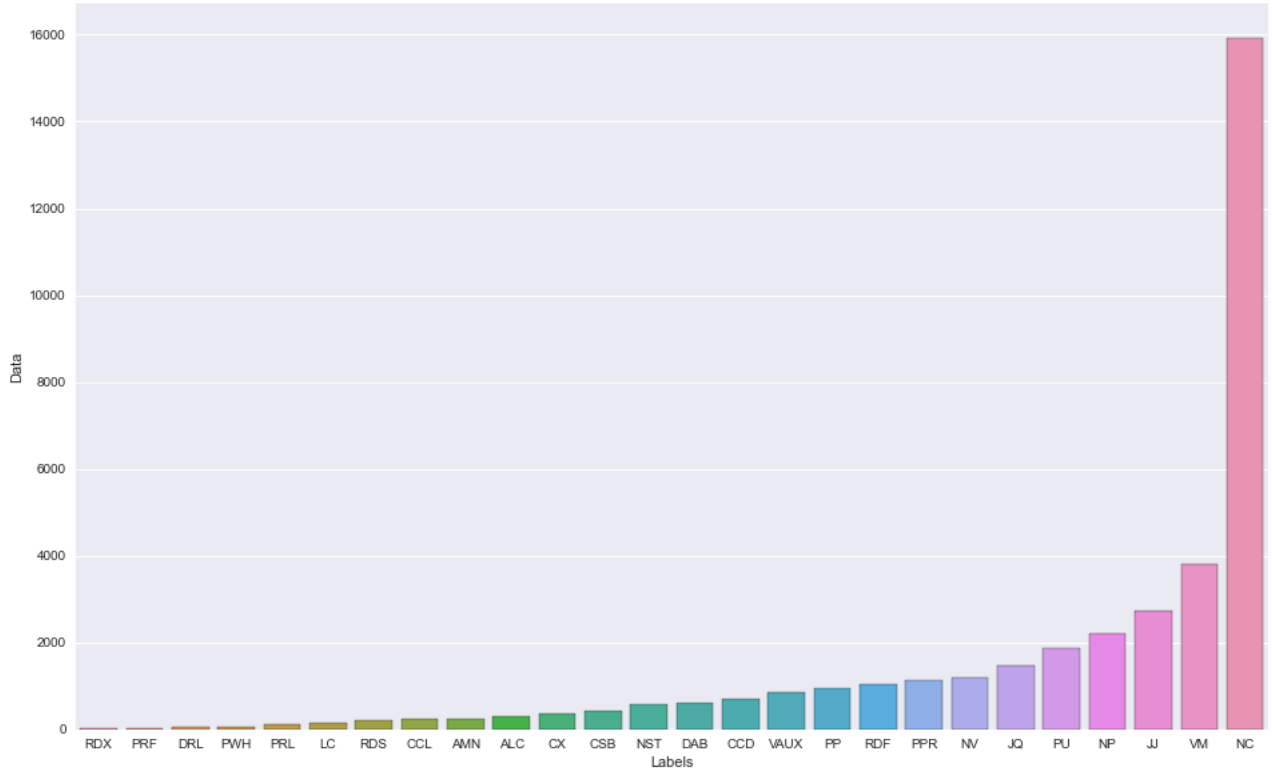

Figure 5.5: Real News Parts of Speech Tagging

5.1.5 Adjectives, Adverbs, Hedging Words and Personal Pronoun Tagging

From our analysis we get the results that in Covid-19 related fake news, there are 0.98% Comparatives, 0.022% Superlatives, 1.32% Personal Pronouns, 0.21% Hedging Words and 0.78% Adverbs.

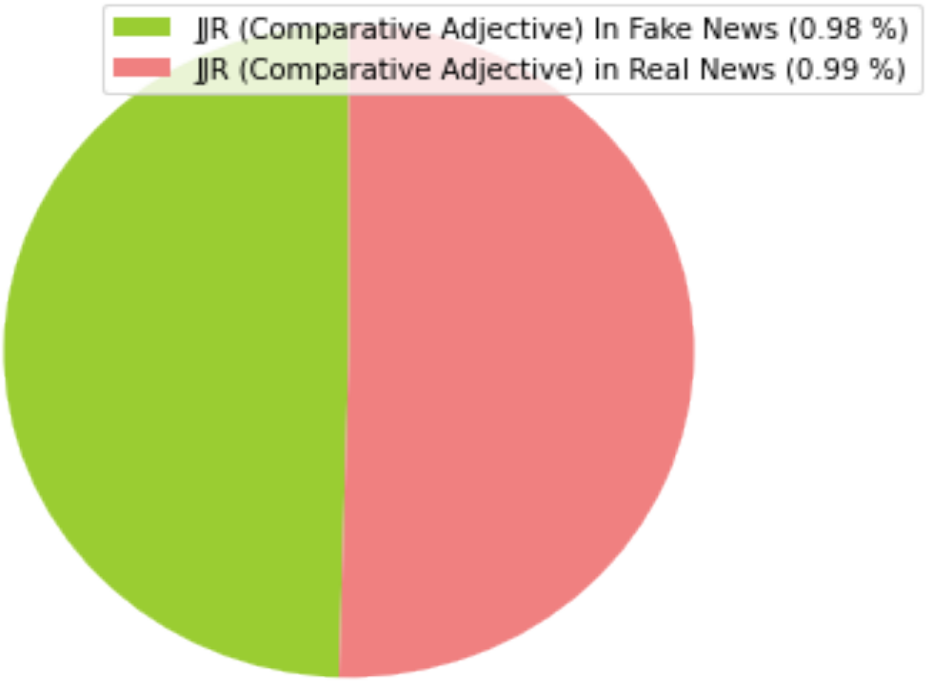

Figure 5.6: Comparative Adjectives

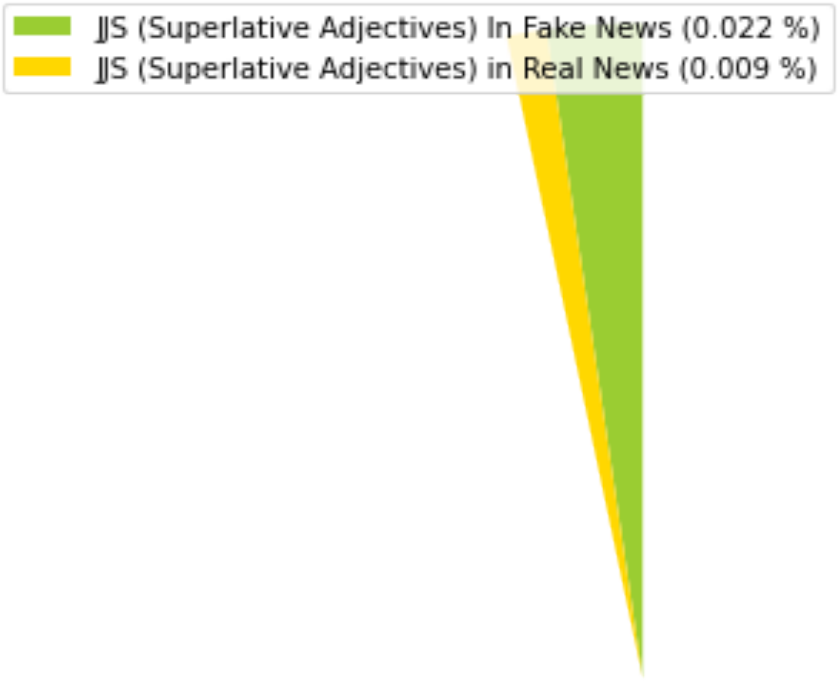

Figure 5.7: Superlative Adjectives

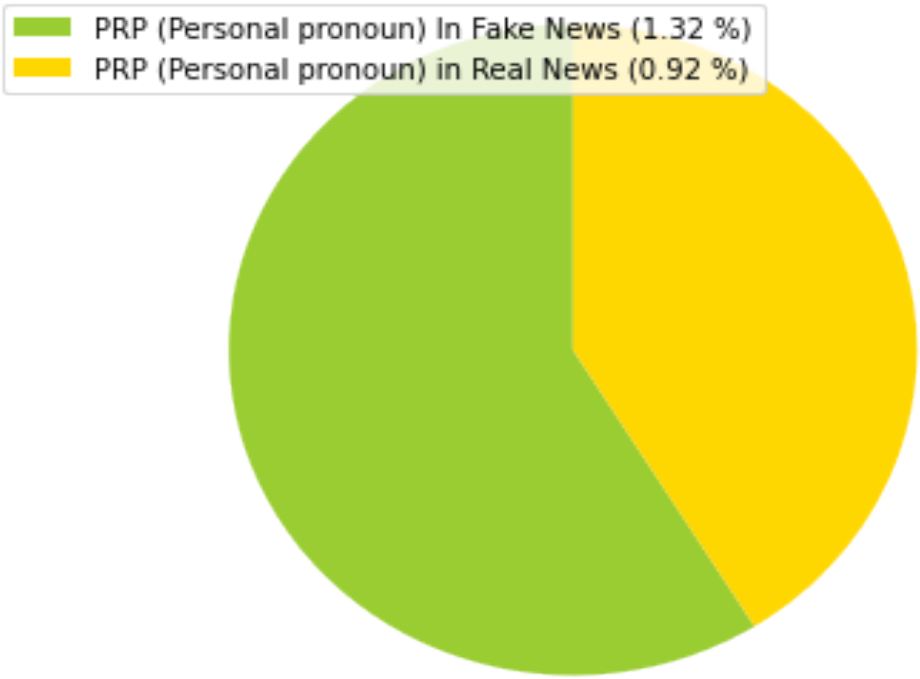

Figure 5.8: Personal Pronoun

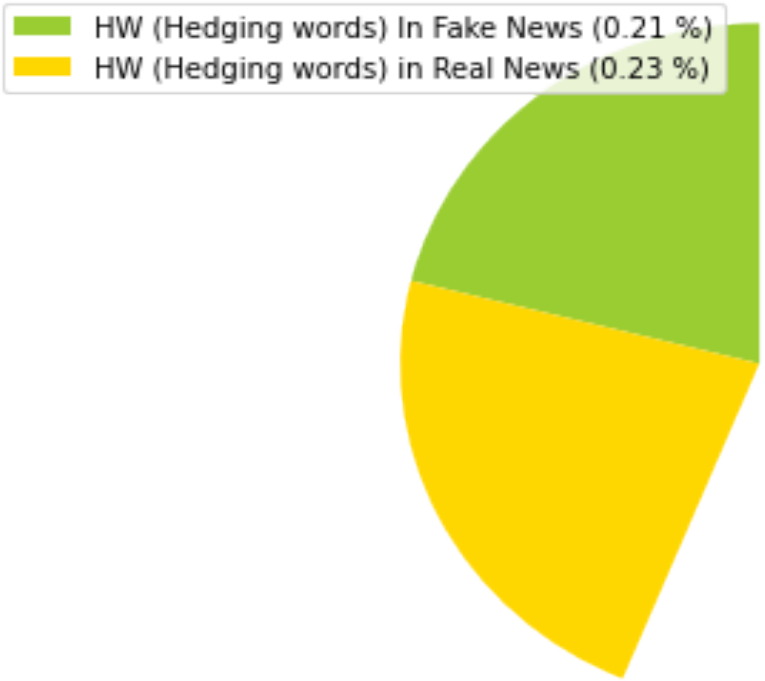

Figure 5.9: Hedging words

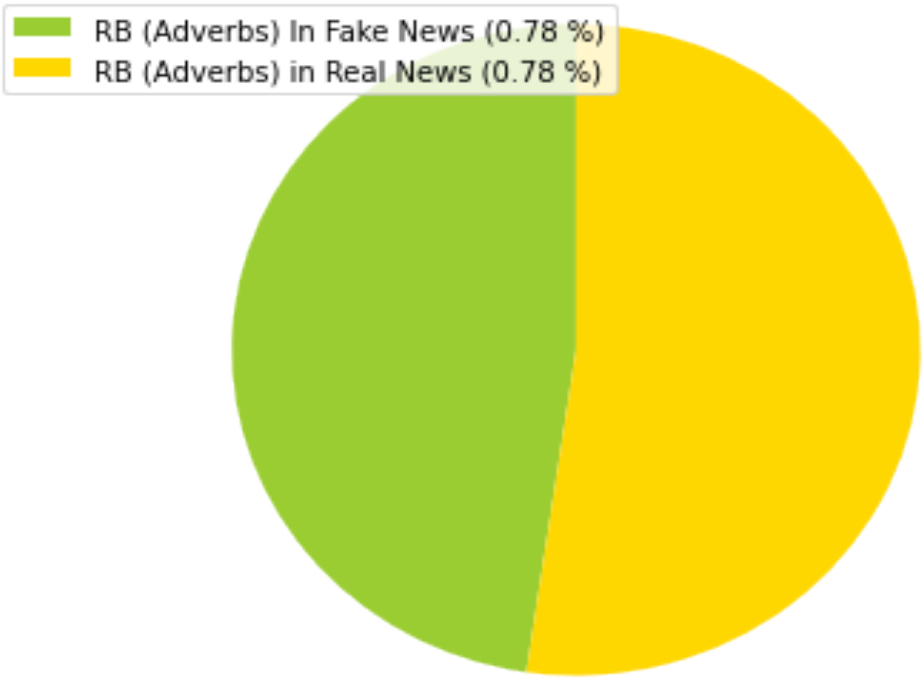

Figure 5.10: Adverbs

In Covid-19 real news, there are 0.99% Comparatives, 0.009% Superlatives, 0.92% Personal Pronouns, 0.23% Hedging Words and 0.78% Adverbs.

5.1.6 Named Entity Recognition (NER)

For NER, in fake news the percentage of Objects is 91.7% while in real news it is 89.4%. The rest of the named entities also have pretty much the same percentage for both datasets.

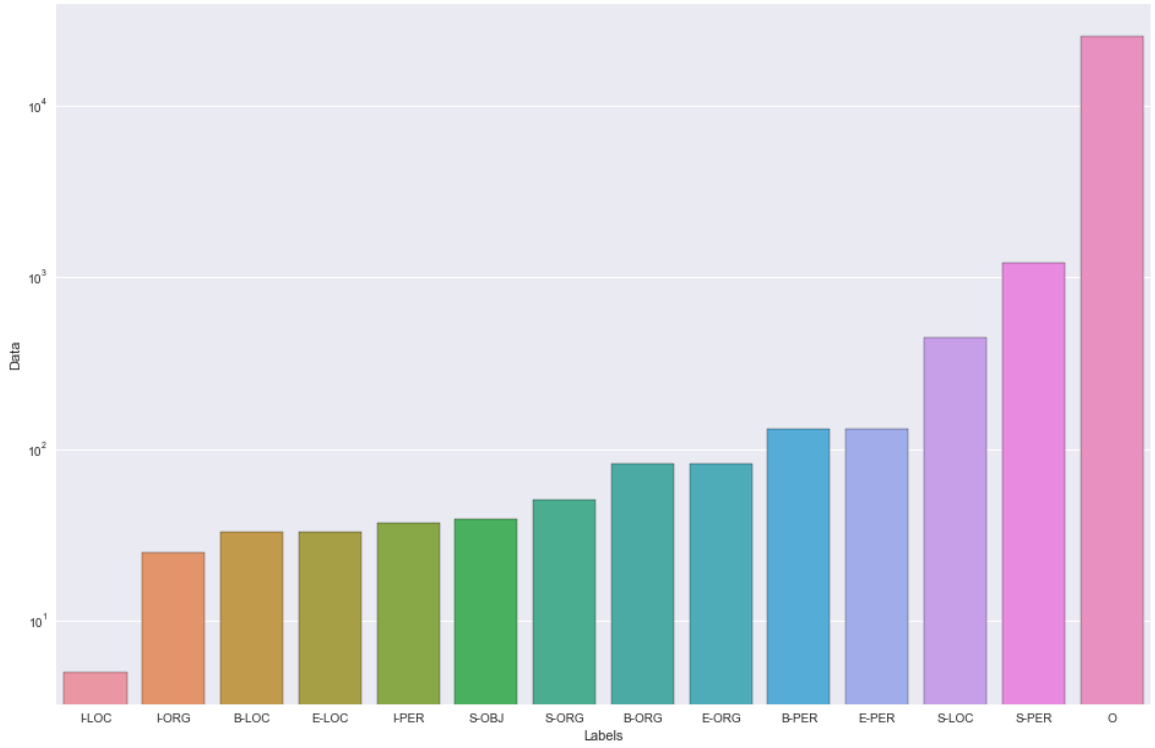

Figure 5.11: Named Entity Recognition for Fake News

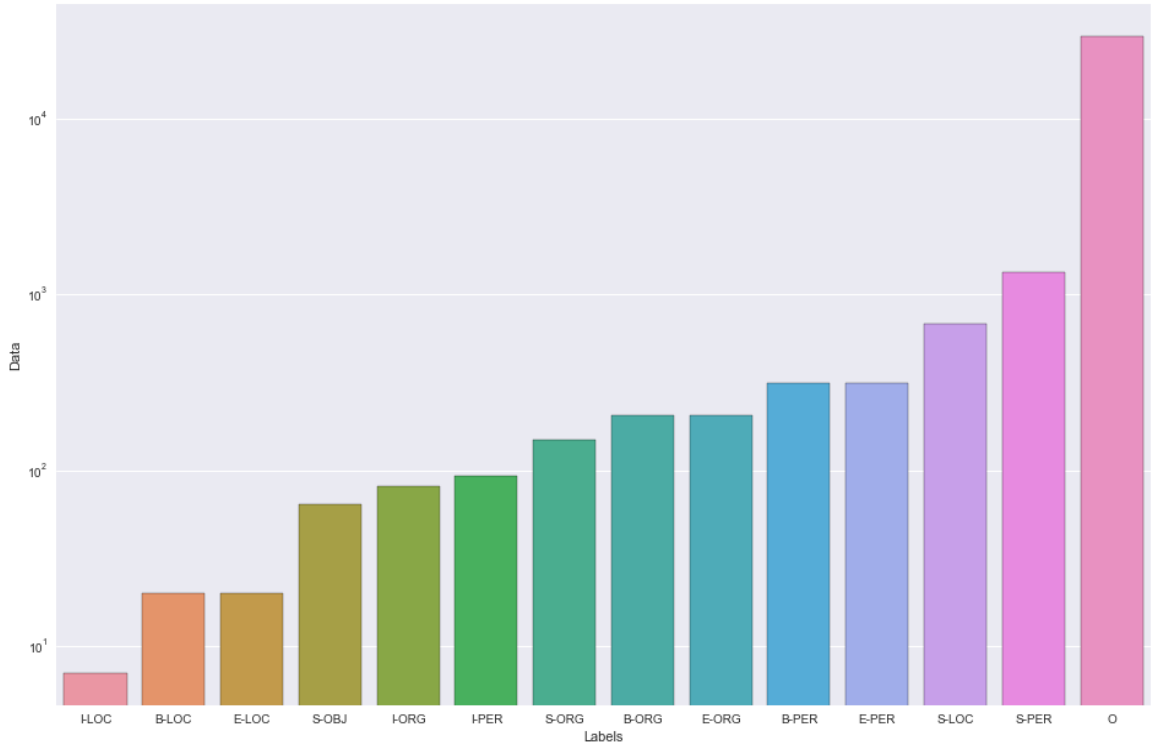

Figure 5.12: Named Entity Recognition for Real News

5.1.7 Unconventional Use of Punctuation

Three types of puctuations marks were searched for in Fake and real news. They are '!', '?' and '.'. Our finding show that in fake news headlines, the number of '!' is 20, '?' is 9 and '.' is 4 while for real news headlines the number of '!' is 3, '?' is 10 and '.' is 2.

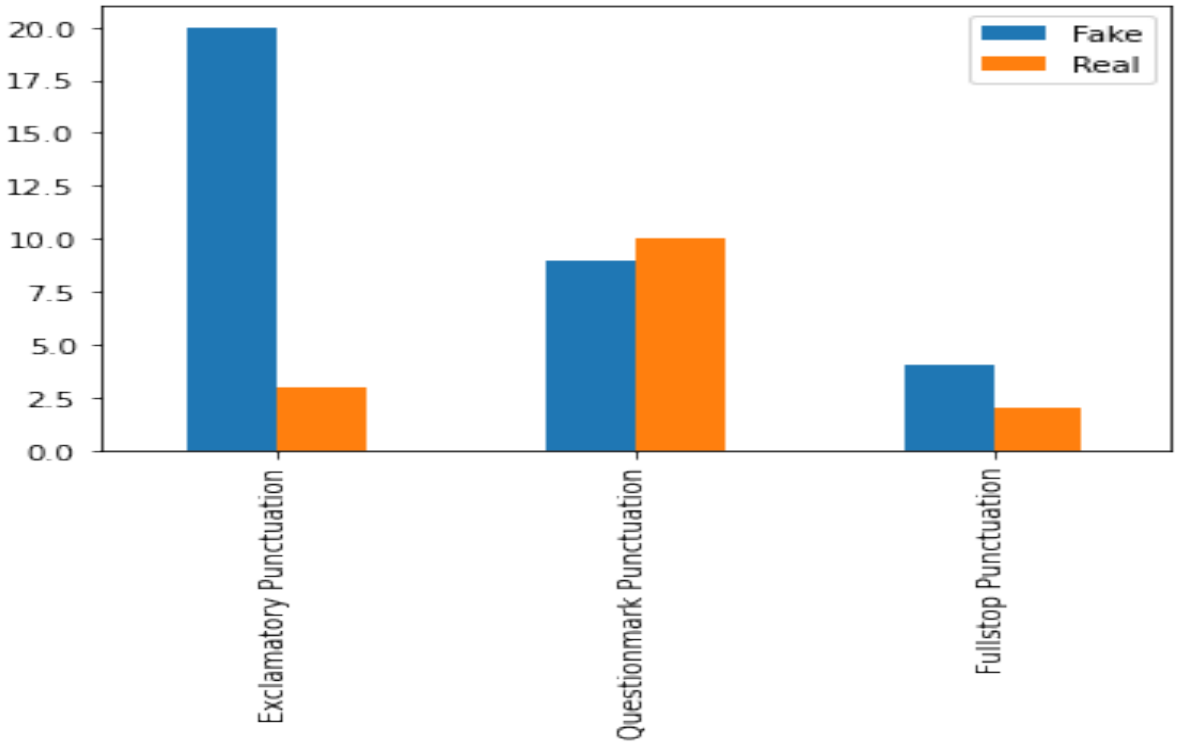

Figure 5.13: Punctuation mark in Headlines

For fake news content, the number of '!' is 30, '?' is 50 and '.' is 10. For real news content, their numbers are 3, 40 and 95 respectively.

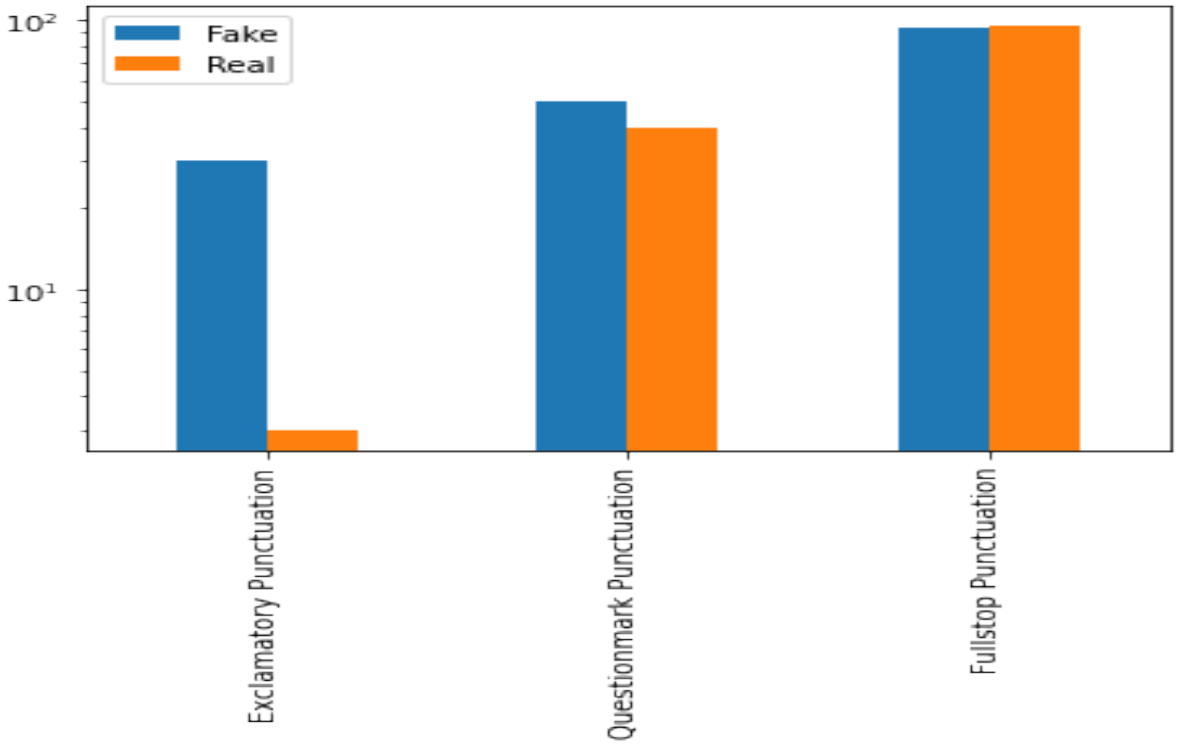

Figure 5.14: Punctuation mark in Content

5.1.8 Number of Images in News Articles

The number of images in Covid-19 related fake news is found to be 88 while the number of images in Covid-19 related real news is 124.

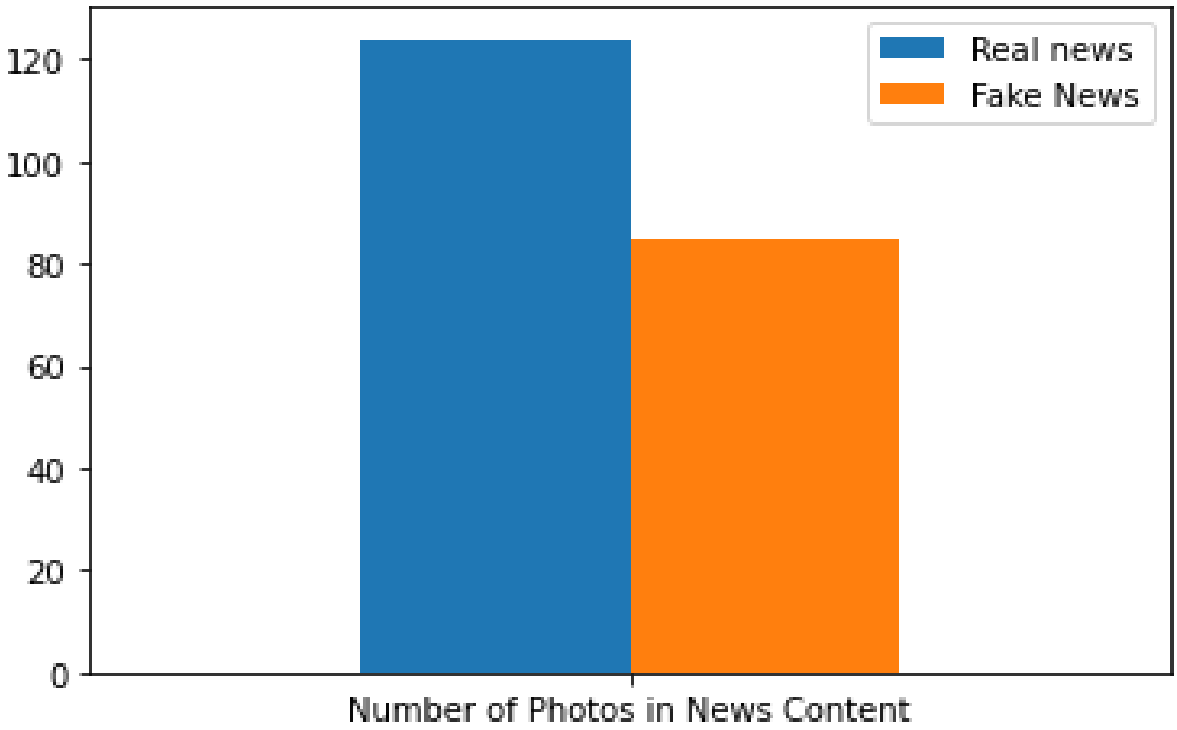

Figure 5.15: Number of Photos in Content

5.1.9 Sentiment Analysis on News Headlines

Positive sentiment in fake news headlines is 35.55% while in real news healines it is 42.22%. Neural sentiment in fake news is 12.22% and in real news is 13.33%. Negative sentiment in fake news is 52.22% while in real news it is 44.44%.

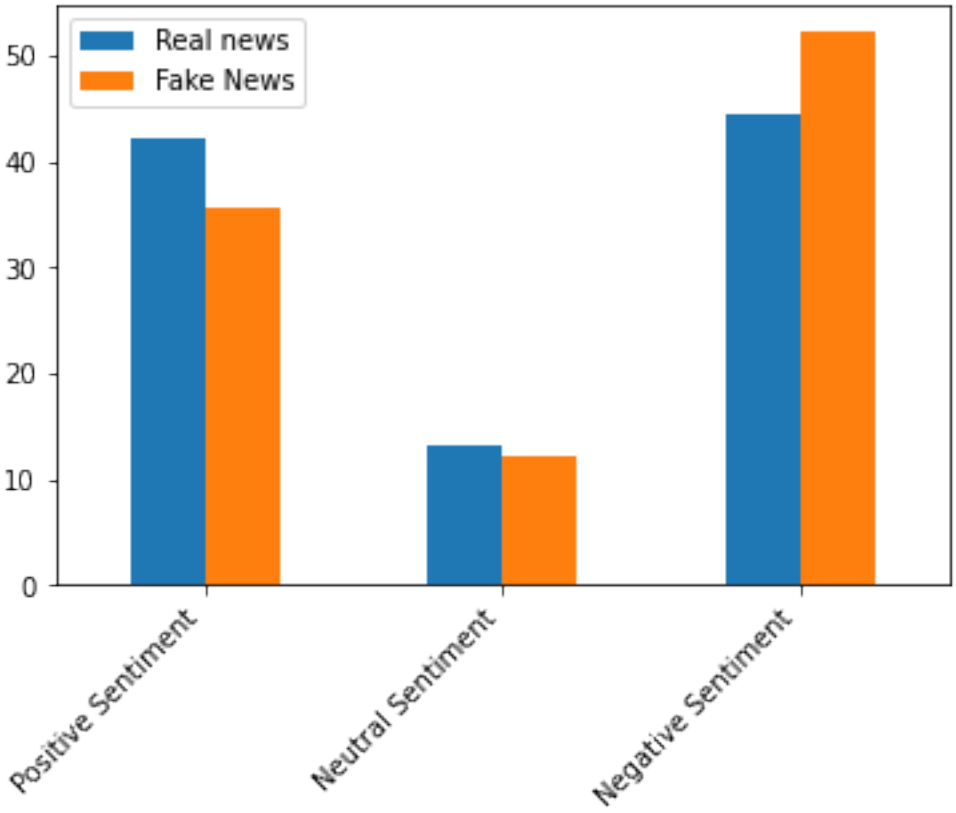

Figure 5.16: Sentiment Analysis Result

5.1.10 Analysis on Online Survey

From the 152 responses that we got from our online survey, the results of the questionnaire show that-

- 79.6% people sometimes read the full news while only 9.9% people read the full news

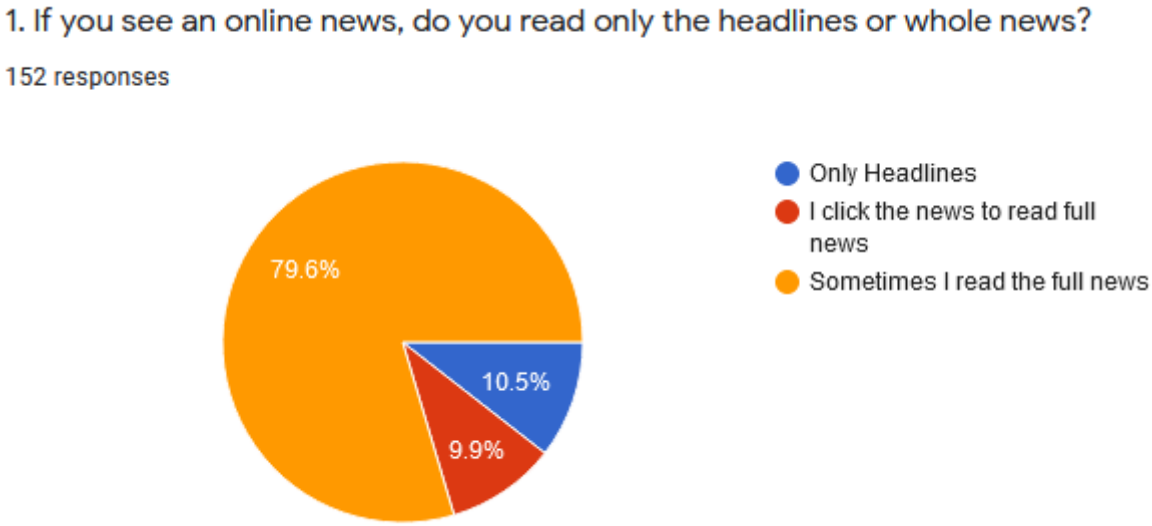

Figure 5.17: Survey Question 1

- The most trusted new protal is Prothom Alo with 66.4% votes, In the second and third position is Bdnews24 and Kaler Kantha with 34.9% and 32.9% votes respectively

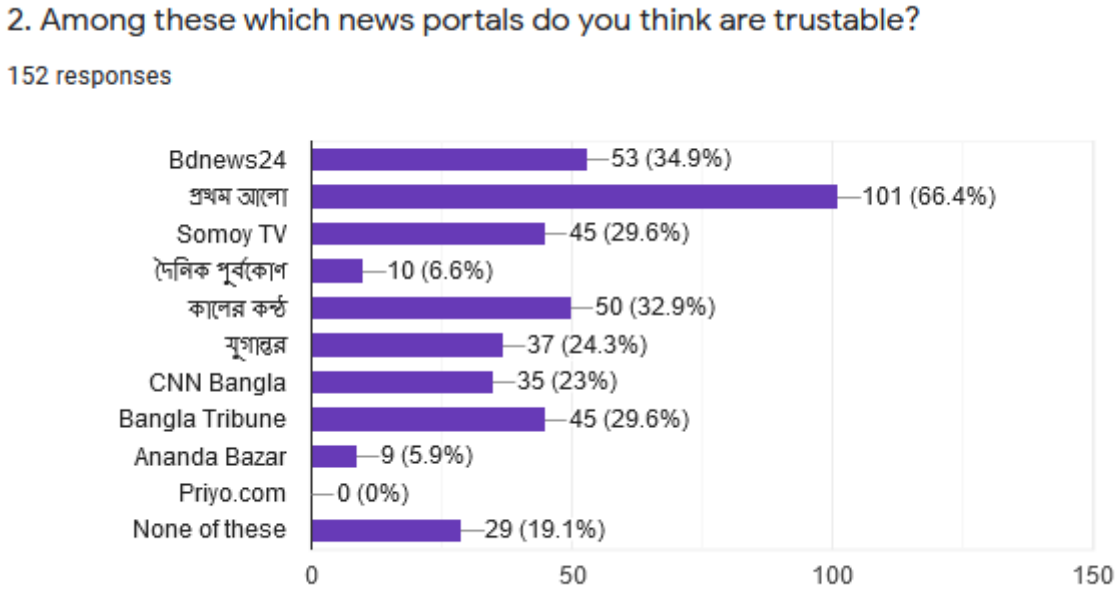

Figure 5.18: Survey Question 2

- 46.1% people have seen clickbaits being shared in the news portals they selected as the most trustworthy, 34.2% have maybe seen it and only 19.7% have never seen clickbaits

being shared in these news portals

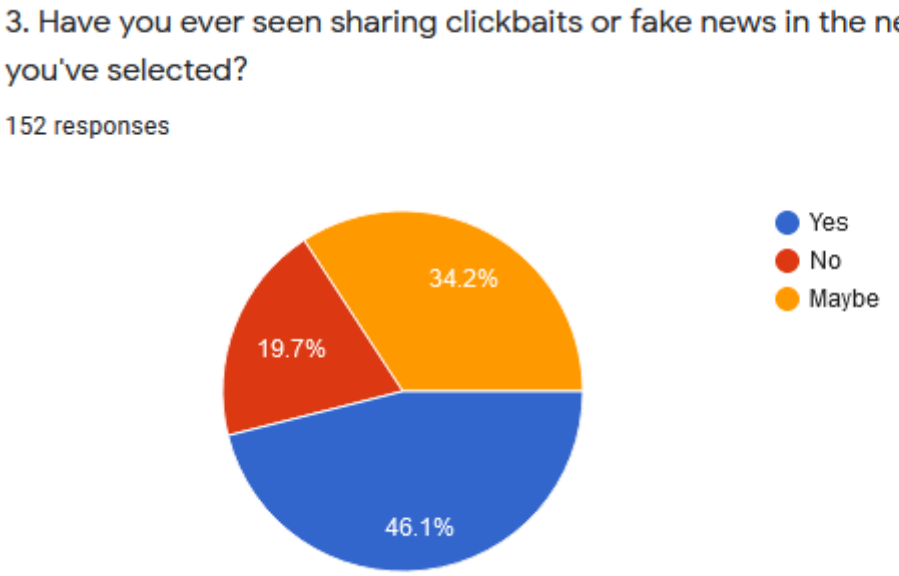

Figure 5.19: Survey Question 3

- 25% people have voted to have shared fake news at least once in their life while 58.6% people have voted to have never shared a single fake news

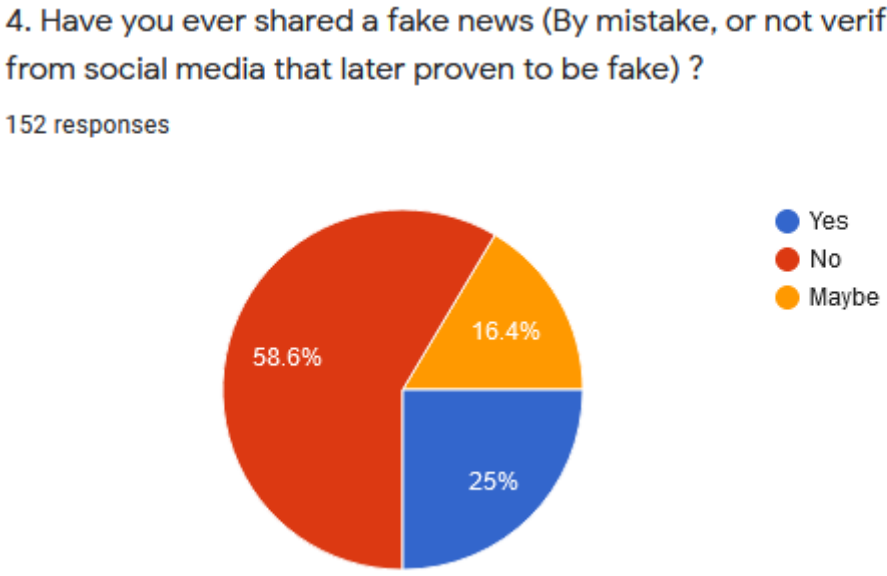

Figure 5.20: Survey Question 4

- 73.7% people have responded that they verify news before sharing, 21.1% people are not

sure if they verify news before sharing or not and only 5.3% people have responded to saying they never verify news before sharing

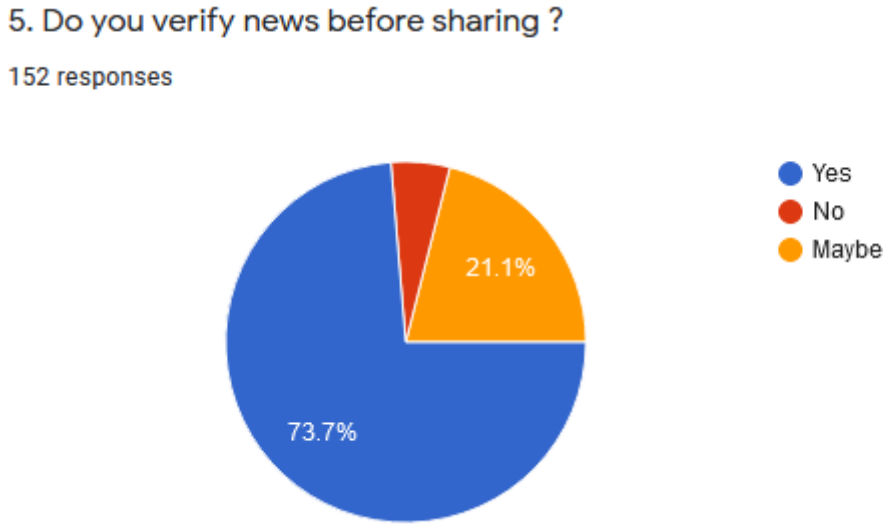

Figure 5.21: Survey Question 5

5.2 Findings

After text analysis and from our observational data the key findings of this study are -

1. Majority of the Covid-19 related fake news are either False or Rumor
2. Majority of the fake news content regarding Covid-19 is related to either Medicine or Prophecy regarding Covid-19 while in real news the topics are Covid-19 Vaccine or Knowledge related.
3. Fake news has longer titles than real news but shorter content.
4. After using our own list of Adjectives, Adverbs, Hedging Words and Personal Pronouns we found that Fake news articles use 2.4% more Superlatives than real news and 1.4% more Personal Pronouns than real news. It indicates that fake news has less substantial information than real news.
5. We found that fake news uses 8% more Exclamatory Sign (!) than real news and almost 50% of the ! in our fake news dataset were in the Headlines of the news. We found 50

Exclamatory Signs (!) in our fake news data set where 20 of them were in the headline but in the real news data set, we only found 6 Exclamatory signs (!).

6. According to our fake news data set, although Kaler Kantha and Somoy TV publishes fake news often, 32.9% and 29.6% people voted them to be trsuworthy news portals.

7. Although other studies give the opposite result, our study shows that real news uses more images than fake news. The number is 1.5% more.

8. Sentiment analysis shows that the percentage of negative sentiment is more in fake news while in real news the percentage of positive sentiment and neutral sentiment is greater. Although this sentiment analysis gives us some result, it is not that great as due to the size of our data set.

9. We did not get satisfactory result using Parts of Speech tagger in our data set as it is too small. Hence the percentage of Parts of Speech is almost the same in both fake and real news data.

10. Named Entity Recognition also does not work well with our data set

Chapter 6

Conclusion And Future Works

We started the textual analysis of fake news in the context of Covid-19 beacause fake news circulation online had reached a its peak at that time. Some of the notable findings of our study show that Covid-19 related fake news has short content but long titles compared to real news, it uses unconventional puctuation (Exclamatory Sign) a lot more, has less substantial informarion than real news and is shared due to the media illiteracy of our people.

6.1 Future Plan

The possible future path for this research includes:

- More analysis regarding punctuation of fake news
- More analysis regrading the use of Superlatives, Comparatives and Adverbs in Bangla fake news
- Analysis on the use of stylistic use punctuation in Bangla fake news

Bibliography

[1] Zhou, Xinyi, and Reza Zafarani. "A survey of fake news: Fundamental theories, detection methods, and opportunities." ACM Computing Surveys (CSUR) 53.5 (2020): 1-40.

[2] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 849–857. DOI:https://doi.org/10.1145/3219819.3219903

[3] Monti, F., Frasca, F., Eynard, D., Mannion, D., Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.

[4] Shu, Kai, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection." Proceedings of the twelfth ACM international conference on web search and data mining. 2019.

[5] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[6] Ferrara, E., "What Types of COVID-19 Conspiracies are Populated by Twitter Bots?", arXiv e-prints, 2020.

[7] Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., Liu, Y. (2020). Coronavirus on social media: Analyzing misinformation in Twitter conversations. arXiv preprint arXiv:2003.12309.

[8] Islam, M. S., Sarkar, T., Khan, S. H., Mostofa Kamal, A., Hasan, S. M. M., Kabir, A., Yeasmin, D., Islam, M. A., Amin Chowdhury, K. I., Anwar, K. S., Chughtai, A. A., Seale, H. (2020). COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social

Media Analysis, The American Journal of Tropical Medicine and Hygiene, 103(4), 1621-1629. Retrieved Feb 23, 2021, from https://www.ajtmh.org/view/journals/tpmd/103/4/article-p1621.xml

[9] Mansur, Munirul. Analysis of n-gram based text categorization for bangla in a newspaper corpus. Diss. BRAC University, 2006.

[10] Laato, S., Islam, A. K. M., Islam, M. N., Whelan, E. (2020). Why do people share misinformation during the Covid-19 pandemic?. arXiv preprint arXiv:2004.09600.

[11] Tasnim S, Hossain MM, Mazumder H. Impact of Rumors and Misinformation on COVID-19 in Social Media. J Prev Med Public Health. 2020;53(3):171-174. doi:10.3961/jpmph.20.094

[12] Hossain, M. Z., Rahman, M. A., Islam, M. S., Kar, S. (2020). BanFakeNews: A dataset for detecting fake news in bangla. arXiv preprint arXiv:2004.08789.

[13] Shahi, Gautam Kishore, and Durgesh Nandini. "FakeCovid–A Multilingual Cross-domain Fact Check News Dataset for COVID-19." arXiv preprint arXiv:2006.11343 (2020).

[14] Jang, Yonghun, Chang-Hyeon Park, and Yeong-Seok Seo. "Fake news analysis modeling using quote retweet." Electronics 8.12 (2019): 1377.

[15] Bandyopadhyay, Samir, and SHAWNI DUTTA. "Analysis of Fake News In Social Medias for Four Months during Lockdown in COVID-19." (2020).

[16] Brindha, Ms D., R. Jayaseelan, and S. Kadeswara. "Social media reigned by information or misinformation about COVID-19: a phenomenological study." SSRN Electronic Journal April (2020).

[17] Azim, Syeda Saadia, Arindam Roy, Amitava Aich, and Dipayan Dey. "Fake news in the time of environmental disaster: Preparing framework for COVID-19." (2020).

[18] Al-Zaman, Md. "COVID-19-related Fake News in Social Media." COVID-19-Related Fake News in Social Media (June 30, 2020) (2020).

[19] Chakravorty, Abhishek, and Shruti Sengupta. "MISINFORMATION, FAKE NEWS, AND IDEOLOGICAL STATE APPARATUS: A STUDY OF COMMUNICATION IN THE LIGHT OF COVID-19 PANDEMIC."

[20] Banerjee, Debanjan, and TS Sathyanarayana Rao. "Psychology of misinformation and the media: Insights from the COVID-19 pandemic." Indian Journal of Social Psychiatry 36.5 (2020): 131.

[21] Datta, Ratul. "INFODEMIC WITH MISINFORMATION AND DISINFORMATION IN PANDEMIC COVID-19 SITUATION: A GLOBAL CASE STUDY." IJARIIE, 2020.

[22] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., Choi, Y. (2017, September). Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 conference on empirical methods in natural language processing (pp. 2931-2937).

[23] Fallis, Don. "A functional analysis of disinformation." iConference 2014 Proceedings (2014).

[24] Saurav, Jillur Rahman, Summit Haque, and Farida Chowdhury. "End to End Parts of Speech Tagging and Named Entity Recognition in Bangla Language." 2019 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE, 2019.

[25] Horne, Benjamin, and Sibel Adali. "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017.

[26] Iswara, Agus Ari, and Kadek Agus Bisena. "Manipulation And Persuasion Through Language Features In Fake News." RETORIKA: Jurnal Ilmu Bahasa 6.1 (2020): 26-32.

[27] Vaibhav, Vaibhav, Raghuram Mandyam Annasamy, and Eduard Hovy. "Do sentence interactions matter? leveraging sentence level representations for fake news classification." arXiv preprint arXiv:1910.12203 (2019).

[28] Karimi, Hamid, and Jiliang Tang. "Learning hierarchical discourse-level structure for fake news detection." arXiv preprint arXiv:1903.07389 (2019).

[29] Das, Debopam, and Manfred Stede. "Developing the Bangla RST Discourse Treebank." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[30] Stede, Manfred, and Debopam Das. "Bangla RST Discourse Treebank: Annotation Guidelines."

[31] Data Set For Sentiment Analysis On Bengali News Comments 10.17632/n53xt69gnf.3file-5b54740b-f054-4f92-bb78-08ab3b95b3bd

References