

# Temporal Analysis of Public Sentiment From Opinions Shared in News Contents



Lomat Haider Chowdhury

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of  
*MSc in Computer Science & Engineering*

December 2023

## Declaration

I, Lomat Haider Chowdhury, declare that this thesis titled Temporal Analysis of Public Sentiment From Opinions Shared in News Contents and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for an MSc degree at United International University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at United International University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

---

Date:

---

(Lomat Haider Chowdhury)

## Certificate

I do hereby declare that the research works embodied in this thesis entitled Temporal Analysis of Public Sentiment From Opinions Shared in News Contents is the outcome of an original work carried out by Lomat Haider Chowdhury under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of MSc in Computer Science and Engineering.

Signature:

---

Date:

---

(Dr. Salekul Islam)

Department of Computer Science and Engineering,

United International University,

Dhaka-1212, Bangladesh.

## **Certificate**

I do hereby declare that the research works embodied in this thesis entitled Temporal Analysis of Public Sentiment From Opinions Shared in News Contents is the outcome of an original work carried out by Lomat Haider Chowdhury under my supervision.

I further certify that the dissertation meets the requirements and the standard for the degree of MSc in Computer Science and Engineering.

Signature:

---

Date:

---

(Dr. Swakkhar Shatabda)

Department of Computer Science and Engineering,

United International University,

Dhaka-1212, Bangladesh.

## Abstract

With the advanced technology and state-of-the-art models, natural language processing and sentiment analysis have been a major area for research. In this research, we have included the implementation of temporal analysis along with sentiment analysis models. The work is based on news article data from a Bengali news channel. Viewers share their opinions below the comment section of the corresponding news item. These news and comments have been a great source of information and research. In this work, we first proposed a dataset containing 7,62,678 public comments and replies from 16,016 video news published from 2017 to 2023 from a renowned Bengali news YouTube channel. The dataset is then processed and data labels are annotated using a machine-learning model after repetitive testing. Data temporal analysis have been done later to identify how the data pattern changes over time for news data. The temporal study has been segmented as year-oriented and topic-oriented analysis. The work reveals that public sentiment shifts based on events that trigger people's opinions. Based on some events, temporal pattern changes from 2017 to 2023 both for the news topic and public sentiment. The work identifies that people are more likely to comment on negative news articles compared to others. Some frequent users are found whose sentiment polarity is mostly positive.

The work is wholeheartedly dedicated to my family who have been my constant inspiration and strength, always been there being my moral, emotional, mental, and financial support.

## Published Papers

Work relating to the research presented in this thesis has been published and submitted by the authors in the following peer-reviewed journals:

1. Lomat Haider Chowdhury, Salekul Islam, and Swakkhar Shatabda, “Bengali News and Public Opinion Dataset from YouTube”, Data In Brief (Elsevier), Vol. 52 (2024) 109938, December 2023.
2. Lomat Haider Chowdhury, Salekul Islam, and Swakkhar Shatabda, “Temporal Analysis of Public Sentiment From Opinions Shared in News Contents” (To Submit).

## **Acknowledgements**

I would like to acknowledge the support of my supervisors Dr. Swakkhar Shatabda and Dr. Salekul Islam who made this work possible. Their guidance and advice carried me through all the phases of writing this thesis.

Finally, I convey my gratitude to the Almighty Allah for His guidance and protection, for letting me get over all the difficulties I faced, and for giving me more than what I asked for.

# Contents

|                                                    |                            |
|----------------------------------------------------|----------------------------|
| <b>List of Figures</b>                             | <b><a href="#">xi</a></b>  |
| <b>List of Tables</b>                              | <b><a href="#">xii</a></b> |
| <b>1 Introduction</b>                              | <b><a href="#">1</a></b>   |
| 1.1 Natural Language Processing . . . . .          | <a href="#">1</a>          |
| 1.2 Sentiment Analysis . . . . .                   | <a href="#">2</a>          |
| 1.2.1 Applications of Sentiment Analysis . . . . . | <a href="#">2</a>          |
| 1.3 Temporal Analysis . . . . .                    | <a href="#">3</a>          |
| 1.3.1 Applications of Temporal Analysis . . . . .  | <a href="#">4</a>          |
| 1.4 Motivation . . . . .                           | <a href="#">5</a>          |
| 1.5 Objectives of the Thesis . . . . .             | <a href="#">6</a>          |
| 1.6 Organization of the Thesis . . . . .           | <a href="#">6</a>          |
| <b>2 Related Work</b>                              | <b><a href="#">8</a></b>   |
| 2.1 Preliminaries . . . . .                        | <a href="#">8</a>          |
| 2.1.1 Word Embeddings or Vectorization . . . . .   | <a href="#">8</a>          |
| 2.1.2 Sequential Feature Extraction . . . . .      | <a href="#">8</a>          |
| 2.1.3 Pre-trained Language Models . . . . .        | <a href="#">9</a>          |
| 2.2 Literature Review . . . . .                    | <a href="#">10</a>         |
| 2.3 Summary . . . . .                              | <a href="#">14</a>         |
| <b>3 Dataset Creation</b>                          | <b><a href="#">16</a></b>  |
| 3.1 Data Collection . . . . .                      | <a href="#">16</a>         |
| 3.2 Data Features . . . . .                        | <a href="#">18</a>         |
| 3.3 Data Preprocessing . . . . .                   | <a href="#">20</a>         |
| 3.4 Data Annotation . . . . .                      | <a href="#">21</a>         |

|          |                                   |           |
|----------|-----------------------------------|-----------|
| 3.4.1    | Exemplary Dataset . . . . .       | 21        |
| 3.4.2    | Model Selection . . . . .         | 21        |
| 3.5      | Summary . . . . .                 | 23        |
| <b>4</b> | <b>Temporal Analysis</b>          | <b>24</b> |
| 4.1      | Data Analysis . . . . .           | 24        |
| 4.1.1    | Statistical Analysis . . . . .    | 24        |
| 4.1.2    | Descriptive Analysis . . . . .    | 26        |
| 4.1.3    | Hypothesis Testing . . . . .      | 27        |
| 4.2      | Year-oriented Analysis . . . . .  | 27        |
| 4.3      | Topic-oriented Analysis . . . . . | 32        |
| 4.4      | Others . . . . .                  | 34        |
| 4.5      | Result Analysis . . . . .         | 35        |
| 4.6      | Summary . . . . .                 | 38        |
| <b>5</b> | <b>Conclusion and Future Work</b> | <b>39</b> |
| 5.1      | Conclusion . . . . .              | 39        |
| 5.2      | Future Work . . . . .             | 40        |
|          | <b>Bibliography</b>               | <b>41</b> |

# List of Figures

|      |                                                                   |    |
|------|-------------------------------------------------------------------|----|
| 1.1  | Basic block diagram for sentiment analysis model . . . . .        | 3  |
| 1.2  | Different patterns of data over time . . . . .                    | 4  |
| 2.1  | Working breakdown for proposed methodology . . . . .              | 14 |
| 4.1  | Number of news comments in the dataset from 2017 - 2023 . . . . . | 25 |
| 4.2  | The ratio of comments and views for the data . . . . .            | 26 |
| 4.3  | Correlation between the numerical attributes . . . . .            | 29 |
| 4.4  | Number of news from each category . . . . .                       | 30 |
| 4.5  | Word cloud from the news hashtags . . . . .                       | 30 |
| 4.6  | Year-wise news and comments information . . . . .                 | 31 |
| 4.7  | Sentiment on Padma bridge from 2017 - 2023 . . . . .              | 33 |
| 4.8  | Sentiment on Bangladesh news from 2018 - 2023 . . . . .           | 34 |
| 4.9  | Temporal analysis of sentiments for corresponding news . . . . .  | 35 |
| 4.10 | Ratio of comments between regular and irregular user . . . . .    | 36 |
| 4.11 | Sentiment-based ratio for news and comments . . . . .             | 36 |
| 4.12 | Result of sentiment annotation Models . . . . .                   | 37 |

# List of Tables

|     |                                                               |    |
|-----|---------------------------------------------------------------|----|
| 1.1 | Number of comments for each polarity over the years . . . . . | 5  |
| 3.1 | A Sample of the raw dataset . . . . .                         | 17 |
| 3.2 | A comparison among other relevant datasets . . . . .          | 19 |
| 3.3 | Model accuracy result . . . . .                               | 22 |
| 4.1 | Rate of news published in 2023 . . . . .                      | 27 |
| 4.2 | Category-wise news information . . . . .                      | 28 |

# Chapter 1

## Introduction

The idea of this thesis is to track the public sentiment on news articles and evaluate how those change over time. Sentiment analysis is a branch of natural language processing. Therefore, NLP and different sectors of using NLP are discussed in this chapter followed by a basic discussion of sentiment analysis and methods, the motivation of this work, the scope of the project, and the organization of the thesis work.

### 1.1 Natural Language Processing

Language processing is mainly used to refer to how the human brain uses words and language to express emotion or communicate [1]. The idea of language processing used by the human brain has then been tried to imitate and developed using computers and named Natural Language Processing. Natural Language Processing (NLP) is a merged field of artificial intelligence and linguistics. The idea of NLP evolved during the 1950s. The implementation of NLP was very challenging at the beginning since it asserts complex grammatical, syntactical, and linguistic perfection. The earliest approach of NLP was to machine translation where Russian to English word-for-word translation was one [2]. After several improvements through trial basis, the idea of syntactic Analysis (Backus–Naur form - BNF and Context-Free Grammar - CFG) was included with the translator for the derivation rule [3, 4]. During the 1970s, a new language “Prolog” was created with the goal of NLP application with the concept of top-down parsing [5]. However, eventually, applying natural language processing using hand-written rules became difficult and demanded the use of statistical NLP where machine learning showed its expertise with the availability of data and computational power [6]. From feature-vector and annotated data machine learning algorithms, like decision trees were able to build statistical rules for parsers [7, 8]. In recent years, with the advancement in deep neural networks and representation

learning, NLP has reached its highest peak and is being used in language modeling, medicine, healthcare, etc [9].

Natural language processing is a large sector to work with. The reason NLP tasks have been segmented into different interdisciplinary parts. The sub-parts of NLP are Sentiment Analysis, Chatbots, Question Answering, Optical Character Recognition, Speech Recognition, Social Media Analysis, Text-to-speech conversion, Voice Assistant, Automated Translation, and many more.

Out of all the segments of natural language processing, sentiment analysis is the focus of this work. Some fundamental discussion about sentiment analysis is represented in section 1.2.

## 1.2 Sentiment Analysis

Sentiment Analysis is a branch of natural language processing (NLP) to identify the emotional tone from text analysis. It also refers to the concept of text analysis and linguistics. Sentiment can mostly be categorized into two-way classification (Positive and Negative) and three-way classification (Positive, Negative, and Neutral). It is a widely used process for psychological research to identify the psychological state of a person [10]. According to [11], methods of sentiment analysis are knowledge-based techniques, statistical methods, and hybrid approaches. Knowledge-based sentiment analysis is mainly for social networks where the unambiguous effect of words is classified and stored in contextualized sentiment lexicon [12, 13]. Statistical approaches include machine learning models like support vector machines, bag-of-words, etc. Sentiment ratings are then evaluated by statistical tests like z-test, chi-square test, p-test, etc [14]. Hybrid methods combine the concept of both the knowledge-based approach and the statistical approach [15].

### 1.2.1 Applications of Sentiment Analysis

The applications of sentiment analysis are many. It can be applied to evaluate the customer calls for different products [16], analysis of the customer reviews and responses [17], public sentiment analysis from social media contents [18], improving the quality of healthcare [19], and so on. To implement sentiment analysis modeling, there are several tools and libraries available now. Some of the popular tools that are being used by the researchers are RoBERTa, Natural Language Tool Kit (NLTK), Bidirectional Encoder Representations from Transformers (BETR), DistilBERT, and AFINN Lexicon. These pre-trained language models have later been discussed in section 2.1.3. Figure 1.1 represents the basic phases of sentiment analysis and its scope from the collected sentiments mainly from different social platforms.

The goal of this work is to apply sentiment analysis models to news article data to identify the changes in public opinion. The task can be broken down into two complementary tasks: Sentiment analysis and sentiment tracking.

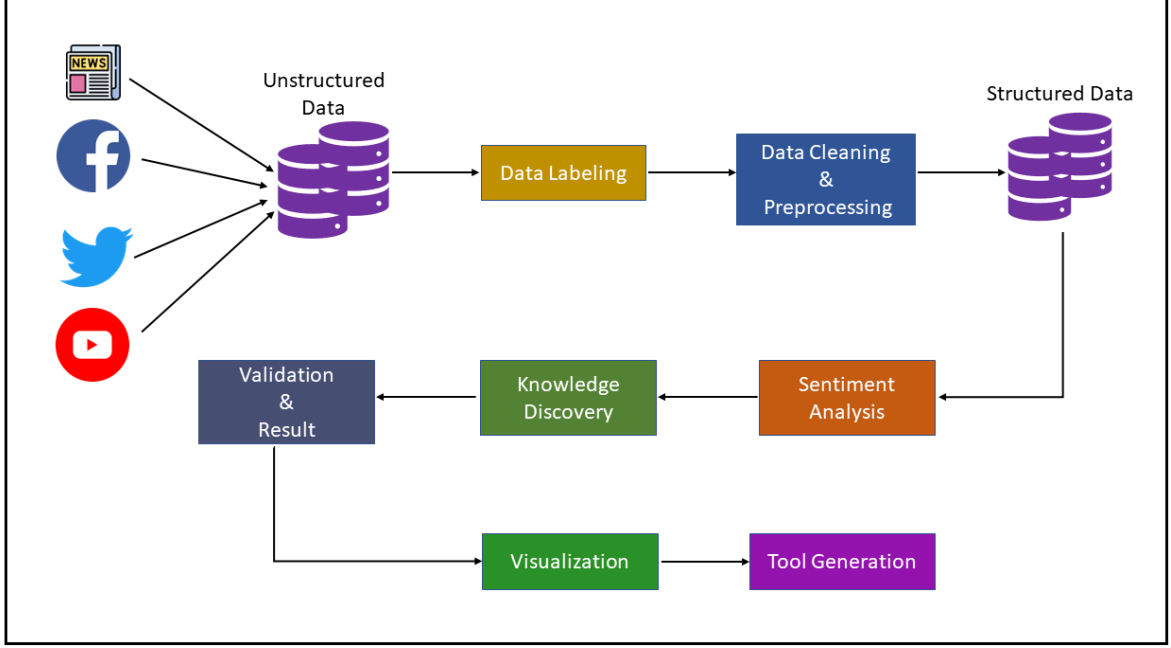


**Figure 1.1:** Basic block diagram for sentiment analysis model

### 1.3 Temporal Analysis

Temporal analysis refers to the study of how patterns, behaviors, characteristics, and trends of data change over time [20]. It is also referred to as time-series analysis because of its time-based mapping. Figure 1.2 represents three different types of time-series plots over a time T1-T8. The behavior of data with time can be positive, thus, upward in direction, it can be downward or even may show a mixed pattern. In NLP, temporal analysis can be used to identify how people's viewpoints change with time.

To further address the temporal pattern of data, we are considering comments from all the published Bengali news in our dataset. Table 1.1 represents the number of comments from 2017 to 2023 for each sentiment polarity. From the table, we can see that the number of neutral comments is always smaller compared to positive and negative comments. Whereas, positive and negative comment varies. There are years where positive comments are higher in number.

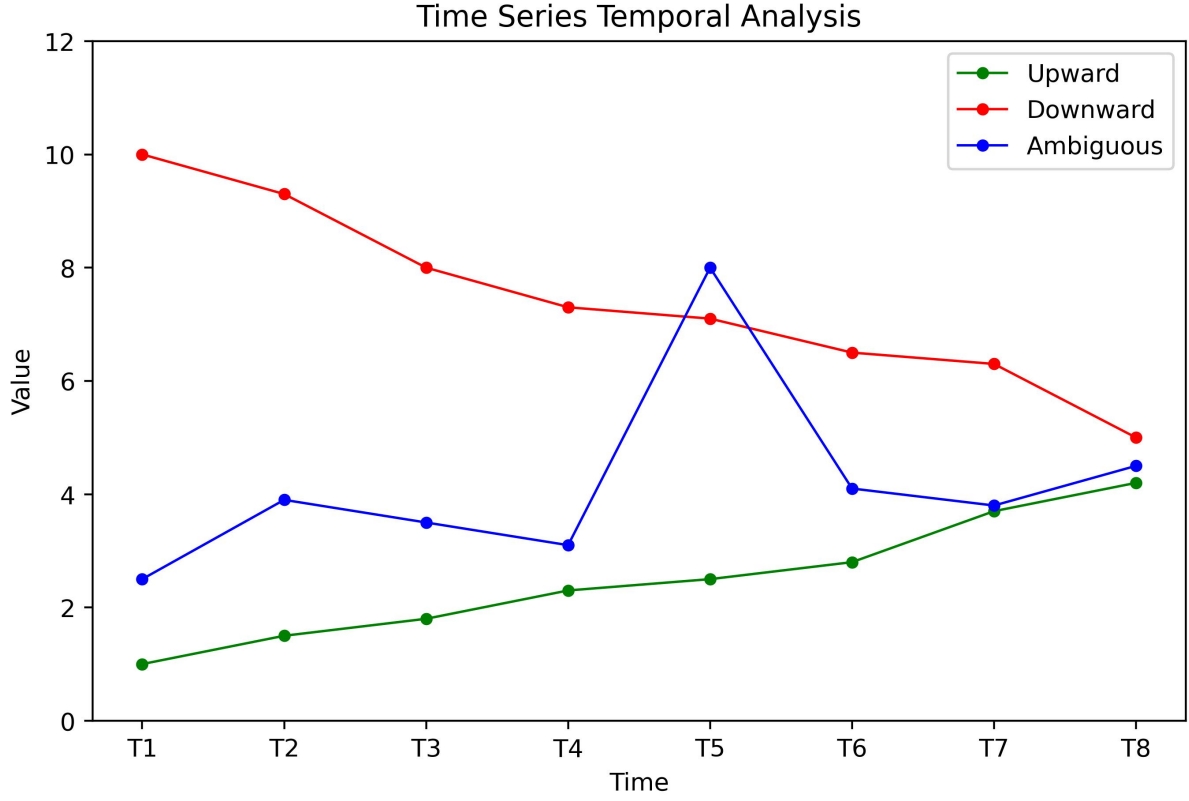


**Figure 1.2:** Different patterns of data over time

In others, negative comments are higher. There must be some events that made this shift in sentiment.

### 1.3.1 Applications of Temporal Analysis

There are vast applications of temporal analysis. It can be used in business to analyze market trends, predict possible changes, and many more [21]. Environmental study and climate change is another branch where spatiotemporal analysis is used widely to identify long-term climate patterns [22]. In healthcare, temporal analysis can be applied to study the trend of a disease, how the treatment for any particular disease evolved over time, and how the patients are reacting to the updated treatment and technologies [23]. One vast application of temporal analysis is in social media platforms. Using the time-series models, social media platforms can determine the most active time of their users, and thus, they can include marketing and advertisement. Also, by following the sentiments of the users on social media, they can update their features and include new ideas to enhance their performance. Over time, social media platforms can determine whether the behavior of their users is in a churn or no-churn trend. If more people

| Year | Neutral | Positive | Negative |
|------|---------|----------|----------|
| 2017 | 650     | 917      | 878      |
| 2018 | 1,619   | 2,219    | 2,087    |
| 2019 | 4,420   | 6,056    | 6,267    |
| 2020 | 9,141   | 13,427   | 13,075   |
| 2021 | 27,977  | 39,388   | 39,556   |
| 2022 | 59,034  | 81,036   | 78,580   |
| 2023 | 100,816 | 144,906  | 130,629  |

**Table 1.1:** Number of comments for each polarity over the years

are engaging in their platform, they should work on how to keep them along. If the customer are being inactive or leaving their platform, they should impose new policies to return them.

## 1.4 Motivation

With the advancement of machine learning and natural language processing, public sentiment is being observed and corresponding actions are taken by the stakeholders. For example, Facebook observes the activity of the users and correspondingly suggests relevant pages and advertisements [24]. Several research have also been done in the past years on analyzing public sentiment based on the posts they share on Facebook or Twitter [25–27]. The outcome of the research helped the psychologist to interpret human mental health, especially during the COVID-19 period [28, 29]. In recent years, sentiment analysis has been a widely popular research topic in natural language processing. Not only in mental health analysis or social media recommendation, but public sentiment can also be studied regarding election [30], vaccine review [31], news articles sentiment [32], and so on.

Some recent research and their outcomes are represented in section 2.2 where most of them are mainly based on social media context. We have found a research gap and scope in sentiment analysis in news article data mostly in Bengali. All the recent research and investigation intrigued our work on this topic. The work was started to combine the study of the sentiments and identify the pattern using time series analysis. In this thesis, we have annotated the public sentiment on different news articles and estimated how the sentiment evolves over time. Tracking opinions over time can be useful for sentiment prediction and identifying the possible reasons for a sentiment change.

## **1.5 Objectives of the Thesis**

The Internet with mobile phones and computers has made our life much easier. Social platforms like Facebook, Twitter, YouTube, etc. have been great sources of information. People are simultaneously the producer of information and the consumer of it. Before the evolution of the internet, Newspapers used to be the most common and easiest source of information. People all over the world used to read newspapers every day to get informed about what's happening at home and abroad. But people nowadays prefer reading news online. News publishing sources have updated their publishing policies accordingly. Numerous online portals share the news from home and abroad over the internet. Additionally, most news media have their own YouTube channels and Facebook pages where they share important news text, graphics, and videos. People who read the articles or watch the graphics and video materials share their opinions as comments below the corresponding media item. Public opinion over different categories is also different. The opinions that the public shares can have positive, negative, or neutral vibes so as the content. Moreover, the news items were also tagged using different keywords based on the criteria. Public opinion towards a specific topic changes over time. We can segment the opinions into two-way or three-way classifications. In two-way classification, the comments can be segmented either as positive or negative. And in a three-way classification, it can be positive, negative, or neutral. Furthermore, we can analyze the aspect-based sentiments of these comments. If we can gather these opinions and examine the viewpoints, it can be a significant source of feedback for the government, policymakers, and everyone. Analyzing public opinion and how it evolves is an interesting part of sentiment analysis.

In this work, we have prepared a dataset from a news portal along with the comments shared corresponding to the news. Then, we implemented some sentiment analysis models to determine the polarity of the comments and news. Alongside, we analyzed how sentiment changes over a while. There are points where we have found changes in the sentiment with time. We identified the factors that trigger this shift in polarity. These factors can be analyzed later to predict public views over a particular topic.

## **1.6 Organization of the Thesis**

This book represents the background study of the work, elementary discussion about natural language processing and sentiment analysis, the idea of the thesis, the motivation and objectives of this work, other related research, the working methodology maintained throughout the task, the scope of the work, result analysis and visualization, challenges, and limitation that we faced

during the thesis, and finally the conclusion and future work. The organization of the book is as follows:

**Chapter 2** discusses some preliminaries and relevant research in this sector.

**Chapter 3** is focused on the discussion about the data collection, challenges, outcomes, and preprocessing of the data along with several sentiment analysis models to annotate the data polarity.

**Chapter 4** is to explain some assumptions and analysis of the data based on time and to incorporate the result analysis.

**Chapter 5** concludes the study and describes the plan to extend the work.

## Chapter 2

# Related Work

With advanced machine learning and a large amount of data, research interest and scope in the particular sector of sentiment analysis have grown over the past years. Before starting the implementation of our thesis, we had to perform a background study on available research publications and articles with similar contexts. This chapter sums them up in short.

### 2.1 Preliminaries

The basic idea and applications of sentiment and temporal analysis are described in Chapter [1](#). Here, we are going to explain some widely used models to identify sentiment and temporal patterns so that further reading of the article becomes more intuitive.

#### 2.1.1 Word Embeddings or Vectorization

Word embeddings are distributed representations of words allowing words with similar meanings to be closer in the vector space. It has become a fundamental component in natural language processing (NLP) and machine learning applications for text data. Vectorization can also be applied where words are represented as numerical vectors. TF-IDF is one of the widely used vectorization techniques that stands for the term frequency-inverse document frequency. It is commonly used in information retrieval and text mining to evaluate the significance of words in a document. Term Frequency measures how often a term (word) appears in a document. Inverse Document Frequency evaluates the importance of a term across a collection of documents.

#### 2.1.2 Sequential Feature Extraction

For text analysis, language modeling, and machine learning, N-grams are commonly used in NLP and computational linguistics. N-grams are contiguous sequences of  $n$  items from a given

sample of text or speech depending on the context. The value of  $N$  determines the number of items in a sequence. For example, a 1-gram or uni-gram represents a single item in a sequence. In the sentence ‘Sentiment Analysis is required’ the uni-grams would be [‘Sentiment’, ‘Analysis’, ‘is’, ‘required’]. A 3-gram or tri-gram represents a sequence of three adjacent items in a text. It can capture more complex relationships between words. In the sentence ‘Sentiment Analysis is required’ the uni-grams would be [‘Sentiment Analysis is’, ‘Analysis is required’].  $N$ -grams of higher orders can be generated similarly, capturing longer sequences of items.  $N$ -grams are used in language modeling, text analysis, speech recognition, and other NLP tasks.  $N$ -grams provide a simple and effective way to capture the local context in a text, however, it may struggle with capturing long-range dependencies or understanding the semantic meaning of phrases. More advanced models, like transformer-based models such as BERT, have been developed to address these limitations by considering the bidirectional context.

### 2.1.3 Pre-trained Language Models

Apart from the machine learning models, some language models are now widely popular for sentiment analysis. BERT is one of the state-of-the-art language models. BERT stands for Bidirectional Encoder Representations from Transformers which is a pre-trained language model developed by Google. Other than sentiment analysis, BERT is used for several natural language processing (NLP) tasks. BERT is a transformer-based model on large unlabeled text corpora to learn contextualized representations of words using two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM task involves masking some words in a sentence and predicting them based on the context, while the NSP task involves predicting whether two sentences follow each other. As the name incorporates, BERT is bidirectional. That means BERT considers both the left and right context for each word. This bidirectional nature helps intuit more accurate and nuanced meanings of words.

BERT converts the input text into WordPiece tokenized embeddings, breaking down words into subwords or tokens. The last layer of the BERT model generates [CLS] token (classification token) which is fed into a classification layer for predicting the sentiment label (positive, negative, or neutral).

Since the BERT model is large and requires substantial computing power, several variations of BERT have been developed to reduce the computational cost. DistilBERT, DistilBERT, RoBERTa, XLNet, and Electra these all are variations of BERT with a reduced number of parameters and model size. BanglaBERT is also one of the variants specially designed for the Bengali language dataset. BanglaBERT (small) and BanglaBERT (large) are available on two different parameter sets.

Similarly, Azure Sentiment Analysis is also a sentiment analysis tool developed by Microsoft Azure. It uses machine learning algorithms to analyze text and provide sentiment scores and labels. The sentiment scores range from 0 to 1, where 0 is negative and 1 is positive. It gives multilingual support and, thus, is suitable for a variety of languages. It provides both real-time and batch-processing options for large datasets. To use Azure Sentiment Analysis, developers typically make API calls to the service, passing the text to analyze and receive the sentiment scores and labels in return. There are other sentiment analysis models like VADER and AFINN lexicon.

## 2.2 Literature Review

In [33], Park et. al worked on sentiment and temporal analysis of Russian campaign messages in the 2016 election and the manipulation of partisan and racial identities. from Facebook advertising and Twitter using Internet Research Agency's (IRA) strategies. The research considered three aspects of communication propagation that are communication platform, partisan identity, and social identity (African-American identity). The authors observed emotions during the campaign period, Russian trolls' strategies to mask left and right-leaning identities, and emotions representing African-American identity. The study identifies a notable presence of arousal and emotional intensity in IRA campaign messages. It also claims that comparatively Facebook messages hold a positive sentiment while tweets are mostly negative (using mean sentiment scoring). The variance of emotions in tweets is smaller as well. The study identified monthly and daily means of sentiment trend and a negative swing was found until 5 months before the election date which later showed a slight positive trend even after the election. On the other hand, the temporal pattern of emotions on Twitter was below 0 and depicted negative emotions throughout the time. The monthly mean was constantly negative yet, it showed a slight positive trend before the election. Another interesting finding from their study is, that monthly emotional variance on Facebook (12.25) is greater compared to Twitter (1.73). The study also identified a notable difference in left and right troll messages on both platforms. The number of left-troll messages (less positive) seems dominant in Facebook whereas right-troll tweets are twice of left-troll tweets (more positive). According to the research, the left-troll messages probably discouraged people from voting and African-American identities seem to show more negative emotions, thus, race may act as a strong form of manipulation.

In [34], Hu et. al proposed Geo-Tweets data from the year 2016 initiated by Harvard Center for Geographic Analysis (CGA). The temporal analysis for the positive, negative, and neutral sentiments is done on a monthly, weekly, and hourly basis. To identify the spatial clusters,

the Local Indicators of Spatial Association (LISA) is employed. According to the findings, during weekend events and F&F gatherings, users are more likely to post positive tweets. From the spatial context, the study finds that people in the western US tend to share pictures of good moments on Twitter compared to other regions. To create a workflow for reproducing and extending the data, the Alteryx platform. For sentiment analysis and scoring, the Valence Aware Dictionary and Sentiment Reasoner (VADER) model is used which is a rule-based tool requiring no training data and providing positive ( $score \geq 0.05$ ), neutral ( $-0.05 < score < 0.05$ ), negative ( $score \leq -0.05$ ), and compound scores (average). To identify the text pattern in the context, the LDA model is used to generate a summary of topics as a discrete probability of distribution over words of each topic. Finally, the study shows monthly, weekly, and hourly Geo-tweets sentiments and the top 30 most salient terms. A weekday distribution of positive sentiment topics is also depicted.

In [35], authors proposed a Graph-based approach for sentiment analysis to detect early cyberbullying on Twitter. At first, keywords are identified using data mining tools Tweepy and Textblob. These keywords are formed as nodes in a graph. The weight of edges is determined by the average polarity of two corresponding keywords in a specific time window.

Paper [36] presented a systematic review and meta-analysis (PRISMA) from the studies on sentiment analysis in research and academia. A total of 41 research articles were reviewed and discussed their objectives, domain, SM platform used, dataset, techniques, and challenges. According to the study, 11 out of 41 research are conducted in the political context. From the platform study, the authors claimed that Twitter focuses on a wide range of subjects including social problems, marketing issues, etc. from different views. Most of the enlisted research used the t-test, ANOVA test, Chi-square test, and Wilcoxon test to determine the significance of sentiment. For sentiment modeling Naïve Bayes, Multinomial Naïve Bayes, Decision Tree, SVM, and Neural Networks were mostly used.

The study [37] is based on a Chinese dataset from the platform Weibo and detects the temporal pattern in sentiment and emotional fluctuations during COVID-19. The dataset is collected from 31 December 2019 to 31 April 2020. Sentiment rating was generated using SnowNLP, a Chinese language-based NLP tool for sentiment analysis. Posts are clustered into 5-time groups for a single day - early morning, morning, afternoon, evening, and night. The study focused on the emotions of surprise, depression, anger, and fear. According to the article, the emotions of fear and anger were more prevalent during the morning and afternoon, and the rest were prevalent during the afternoon and night. They used “groupby” method to calculate the frequencies of emotions at a particular time of the day. The authors created a heatmap based on the geolocation of the users and demonstrated a correlation pattern between the number of Weibo

posts published by users and their distance to major hospitals. To address the temporal study, the authors found an increase in all emotions during mid-January 2020 when the confirmation of human-to-human transmission of the virus was stated. A smooth pattern in the emotions was addressed till the first week of April 2020. Fear was the prevalent emotion during the end of January. A peak in depressive emotion was found just after Chinese New Year's Day and was at its highest around Tomb Sweeping Day. A peak of emotion in surprise was identified on January 30, 2020, when the last province of China declared the outbreak of COVID there. The emotion of anger reached its highest peak when the declaration for isolation came.

Lorentz et al. in [38] discussed how public comments and corresponding sentiment on YouTube videos affect the like proportion of the videos. For the data, the authors used the trending top list of YouTube videos of the UK in 2017, and also, used the data from Twitter for comparative analysis. The study claims comments on Twitter hold a negative tone more, whereas YouTube comments are mostly neutral. According to the methodology, the authors considered neutral comments in several ways for model implementation. The neutral comments were considered to be positive, and negative, and also segmented equally in both groups. They implemented Logistic Regression, Support Vector Machine, Stochastic Gradient Descent Classifier, Complement Naive Bayes, and Multinomial Naive Bayes. Among them, Logistic Regression performed best to identify the video liking proportion. Though the study was thorough enough, the authors claimed that it has a less real-world application. They also suggested that the model can be improved if emoji sentiment is considered.

In [39], the authors used a deep neural network model for the Brazillian Portuguese Language to classify comments on YouTube videos. For the data, they used two video data along with the comments from a renowned journalist and YouTuber of Brazil, Joice Hasselmann. The focus of the work is to analyze the sentiment of viewers regarding the video quality, the relevancy of the video, and how well the video is explained. The authors segmented the methodology as Lexical preprocessing (to remove unnecessary punctuations), word dictionary extraction (most frequent words only), vectorization, and deep neural network for automatic feature learning. A 6-layer neural network is used for the model. The first embedding layer removes unnecessary information and represents a compact dimensionality of the data. The next two layers are the convolution layers with 64 and 16 features respectively. After that, a dense layer is implemented with 50 neurons and finally, a dropout layer is used to overcome the overfitting problem. They used the 'tanh' activation function for the output later that outputs the value from -1 to 1 where [-1, 0) represents negative sentiment, 0 represents neutral sentiment, and (0, 1] represents positive sentiment.

In [40] authors prepared a news corpus on Arabic news and its associated sentiment tracking system to monitor the sentiments towards news entities in the Arab world. The researchers collected the data from Twitter along with the geographical location of the tweets. For data labeling, they used semi-automated and manually annotated datasets. The tweets are labeled as a positive-negative-neutral, geo-location label, organization, list of hashtags, etc. Four machine learning classifiers were used including, Logistic Regression (LR), Multilayer Perceptron (MLP), Naïve Bayes (NB), and Support Vector Machine (SVM) where semi-automated annotation showed a better result for different machine learning models. They generated a tool for semi-automated sentiment annotation. The work then differentiates how sentiment changes based on geographical locations. The research identifies that people from Egypt share more tweets both for positive and negative sentiments compared to the other twenty locations of Arabs.

Paper [41] explored conventional time series analysis methods and their applicability to topic and sentiment trend analysis. For identifying the opinionated terms, the AFINN Lexicon was used. The AFINN lexicon is one of the simplest and most popular lexicons that can be used extensively for sentiment analysis. Using the time series, the authors identify the outliers of the data that cause sentiment change corresponding to some manually detected tweets. They used frequency measures of tweets for a specific sentiment (android lollipop, Merkel, Michelle Obama). Trend (Tt), seasonal (St), and random (Rt) are considered components of time series decomposition. Another paper [42] reports on the design of a sentiment analysis tracking algorithm, a web-based application called the Villanova Analytical Sentiment Tracker (VAST) for Twitter with two main components of tracking and evaluating.

Public health emergencies at the same time are national and global concerns where precautions and actions should be taken as early as possible. In [43], Xiang Ji et al. present an Epidemic Sentiment Monitoring System (ESMOS) that provides tools for visualizing Twitter users, a tool for public health specialists and government decision-makers to gauge the degree of concern (DOC) expressed in the tweets towards different diseases. They used disease keywords to distinguish the tweets. Among the implemented machine learning models are Support Vector Machine (SVM), Naive Bayes (NB), and Multinomial Naïve Bayes (MNB). MNB achieved the best results with a time of 0.18 sec to build the model.

Other research has been done to predict dangerous events from social media posts [44]. According to the research, events that have a common ground are represented as dangerous events. The dangerous events can be segmented as sentiment-based (emotional and psychological attitude), scenarios-based (not previously planned), and action-based events (human involvement).

Sport is a field where sentiment analysis has a lot to contribute. In [45], authors Faruque et al. used Naive Bayes, Support Vector Machine, and Logistic Regression to analyze the sentiment polarity of people regarding Bangladesh Cricket. For data collection, they used Facebook posts and comments from newspaper pages and another ABSA dataset. According to their result, Logistic Regression performed the best to identify the polarity (Positive, Negative, Neutral) of public opinion.

We have found a research gap in sentiment analysis with the news articles data. Most of the works are mainly focused on Twitter or Facebook. To our knowledge, public sentiment tracking on news channel publications and users' comments is a fresh field to work on. Moreover, minimal work has been done on the Bengali language since no proper dataset is available. The work's objective is to analyze the public sentiment on the comments of the news events published on different news portals and identify how people's sentiments change over time on a particular event.

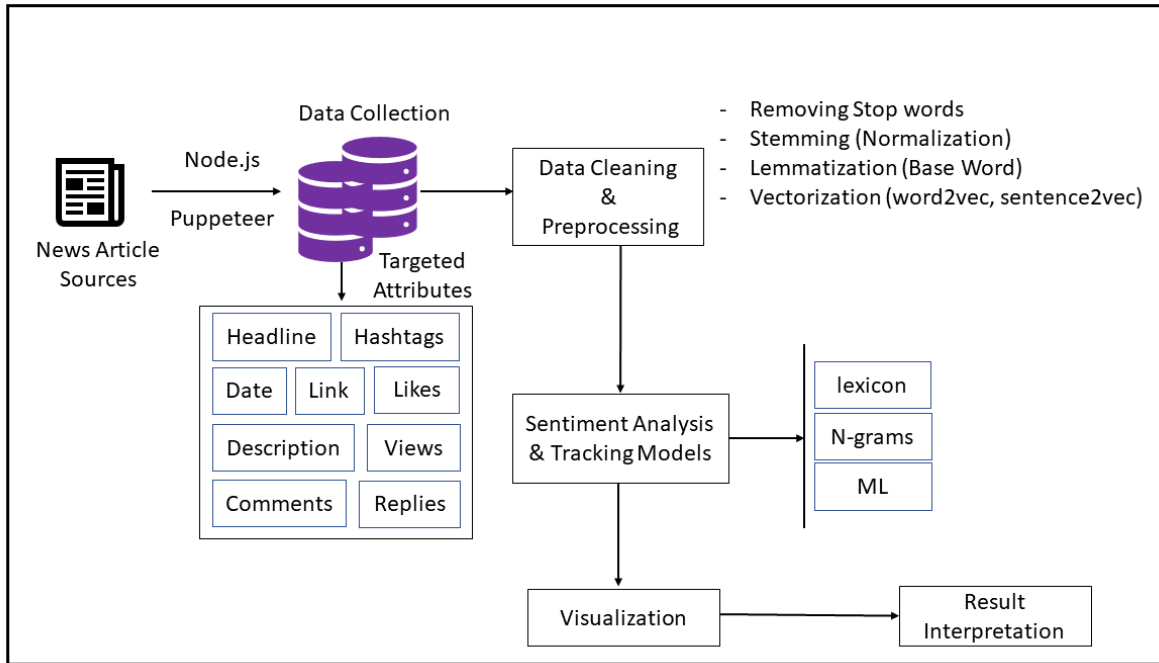


**Figure 2.1:** Working breakdown for proposed methodology

## 2.3 Summary

Though some research works have been done in other languages like English, Chinese, and Arabic, Bengali news article data remain uncovered. To start with, we need to prepare a

dataset. A tentative workflow and methodology of our work are depicted in Figure [2.1](#).

## Chapter 3

# Dataset Creation

After conducting the background study and literature review, we discovered that many minimal works have been done in the context of sentiment and temporal analysis. Moreover, no proper Bengali dataset was available that covers enough period for temporal analysis. Thus, we decided to prepare a dataset focusing on our goal; this data collection was the most challenging part of this task. In this part, we will describe the steps of data collection, the challenges we faced, and how we processed our data to prepare a final dataset.

### 3.1 Data Collection

To understand how native Bengali-speaking individuals feel about national and international news articles, we need a proper dataset that covers a variety of news. With that goal, the data collection process was initiated.

For data collection, initially, we targeted the news portals and websites of Bangladesh. The news agencies shared their collected news in these portals and people can share their opinions below in the comment box. After collecting around 1200 data from these portals, we figured out that the number of comments here is below 40%. This means people are less likely to share their opinions via these news portals or news websites. Therefore, we rethought to change the source of our data. Considering different social media platforms, we decided to collect the data from YouTube since it engages 122 million people daily according to the YouTube Statistics 2023 by Global Media Insights [46].

The source of our data collection is **Somoy TV**, a renowned Bengali YouTube news channel with more than 20 million subscribers [47]. The main reason behind choosing this channel was their highest number of views among Bengali news channels. The data is gathered from YouTube using NodeJS (v18.12.1). The Chrome browser is controlled for URL retrieval using

### 3.1 Data Collection

Table 3.1: A Sample of the raw dataset

| video<br>Url                                | title                                                                                                                             | total<br>Likes | views | date              | hashtags                   | descrip<br>tion                                                         | comments                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|----------------|-------|-------------------|----------------------------|-------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| https://www.youtube.com/watch?v=TUCkdcAn0Q  | ভোলায়<br>তুলার<br>গুদামে<br>ভয়াবহ<br>আগুন<br> <br>Bhola<br> <br>Cot-<br>ton<br>Ware-<br>house<br>  So-<br>moy<br>TV             | 81             | 2,201 | Apr<br>3,<br>2023 | [#bhola,<br>#so-<br>moytv] | আরও<br>বিস্তারিত<br>জানতে<br>ভিজিট<br>করুন:<br>https://www.somoynews.tv | ['cauthor': 'Author A', 'ctime': '28 minutes ago', 'comment': ' কি ভয়াবহ দৃশ্য, আল্লাহ সকলকে হেফাজত করুন আমিন', 'likes': '0 ', 'replies': [], 'cauthor': 'Author B', 'ctime': '2 hours ago', 'comment': 'আল্লাহ তুমি আমাদের এই আগুন নামক ভয়াবহ দুর্যোগ থেকে ক্ষমা করুন ', 'likes': ' 0 ', 'replies': [], 'cauthor': 'Author C', 'ctime': '2 hours ago', 'comment': 'খুব দুঃখ জনক ঘটনা ', 'likes': ' 0 ', 'replies': []]                                                                                                                                 |
| https://www.youtube.com/watch?v=H8UMpRYipFg | ছোট<br>বেলার<br>নেশাটাই<br>এখন<br>পেশা<br>সোহেল<br>রানা'র!<br>  So-<br>hel<br>Rana<br> <br>Bang<br>lade<br>shi<br>Foot-<br>baller | 119            | 5,494 | Aug<br>5,<br>2020 | []                         | []                                                                      | ['cauthor': 'Author D', 'ctime': '1 year ago', 'comment': 'আমি খুব ভালো ভাবে চিনি তাকে', 'likes': ' 0 ', 'replies': [], 'cauthor': 'Author E', 'ctime': '1 year ago', 'comment': 'Nice mama ', 'likes': ' 0 ', 'replies': [], 'cauthor': 'Author F', 'ctime': '2 years ago', 'comment': 'Amar jela , Amar gourab', 'likes': ' 2', 'replies': ['rauthor': 'Author G', 'rtime': '2 years ago', 'rep': 'Manikgonjer kothai bari vai', 'rlikes': ' 0', 'rauthor': 'Author H', 'rtime': '8 months ago', 'rep': 'শিবালয় থানার সাকরাইল গ্রাম', 'rlikes': ' 0']] |

the Puppeteer NodeJS library by Google. We used Puppeteer to automate the browser for web scraping. It enables dynamic interactions with the content and simultaneously retrieves the data. A delay time of 2000ms was used to let the browser scroll down and load the content and comments.

Both sequential and keyword-based searching were used for data collection purposes. The data collection process was initiated in May 2023. A sequential data collection method is used to retrieve all of the news information from 23 May 2023 back to 27 December 2022. A total of 13 features were collected for each data including news URL, title, views, likes, date, hashtags, description, comment author, comment time, comment, reply author, reply time, and reply. The features are described briefly in Section 3.2. Additionally, a keyword-based search was integrated to collect particular categories of news. A total of 19 keywords are used to retrieve the data. The keywords are mentioned in Table 3.2 as “Categories of Data”. Duplicate news entries were not checked during the data collection phase. A total of 22, 757 news data were collected at this point. These data represent news within the timeframe of 2017 to 2023. The raw data is retrieved in a JSON file for further processing. Two samples from the raw dataset are shown in Table 3.1. To ensure the commentator’s privacy, their names have been hidden.

## 3.2 Data Features

We collected a total of 9 features corresponding to each news as described in Figure 2.1. At first, the URL for the corresponding news data is stored as a feature “videoUrl”. The title for the news articles is the second feature stored as “title”. For most of the samples, the video title is represented as the title in Bengali, the title in English, and the channel name, separated by a vertical bar (|). After that, the total number of likes and the total number of views are stored as two different features named “totallikes” and “views” respectively. The date on which the news was published is stored in the “date” feature. While publishing news, the publishers use some hashtags relevant to the news to make the search convenient to the users. These hashtags were also collected as a feature. Publishers can choose to publish news without hashtags. When no hashtags are found for the news an empty square bracket []. There is a description box below each video on YouTube that involves some relevant information regarding the news or publishers along with Copyright details as well. Discarding the copyrights, the rest of the details are stored in the “description” feature. After that, the comments and replies were collected which are the main focus of this dataset. For each comment and reply, four different attributes were stored as “commentAuthor”, “commentTime”, “comment”, “commentLikes”, “replyAuthor”, “replyTime”, “reply”, and “replyLikes”. Here, the author is the person who comments or

**Table 3.2:** A comparison among other relevant datasets

| Specification          | KanCMD [48]                                                                                        | Russia-Ukraine War [49]         | ABSA [50]                                                                             | BAN-ABSA [51]                          | Proposed Dataset [52]                                                                                                                                                                             |
|------------------------|----------------------------------------------------------------------------------------------------|---------------------------------|---------------------------------------------------------------------------------------|----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Number of Comments     | 7,671                                                                                              | 10,861                          | 5,091                                                                                 | 9,009                                  | 7,62,678                                                                                                                                                                                          |
| Platform               | YouTube                                                                                            | YouTube                         | Facebook                                                                              | News Portals                           | YouTube                                                                                                                                                                                           |
| Language               | Kannada, Latin, and English                                                                        | Bengali                         | Bengali                                                                               | Bengali                                | Bengali                                                                                                                                                                                           |
| Categories of Data     | Movie trailers, Trends of banned mobile apps in India, India-China Border, Mahabharat, Transgender | Russia and Ukraine War          | Cricket and Restaurant Review                                                         | Politics, Sports, Religion, and Others | Latest news, Special news, Politics, Bangladesh, World, Crime, Trade, Opinion, Sports, Entertainment, Job, Lifestyle, Education, Election, FIFA, Earthquake, Price hike, Padma bridge, Metro rail |
| Attributes or Features | Comments, Class                                                                                    | Video URL, Comments, Annotation | Cricket: Source, Date, Text, Category, Polarity, Restaurant: Text, Category, Polarity | Comment, Aspect, Polarity              | Video URL, Title, Likes, Views, Publish Date, Hash-tags, Description, Commentator, Comment, Comment Time, Comment Likes, Reply Author, Reply, Reply Time, Reply likes                             |

replies. To ensure the privacy of the commentators, the name of every person who comments and replies is replaced with a unique author count. “commentTime” and “replyTime” represents the time of the statement made from the data collection date. “comment” and “reply” here are the statement made by people regarding the video which usually depicts public’s sentiment and point of views. The number of likes in the statements is stored as “commentLikes” and “replyLikes”. Though there is a dislike option attached to each video, comments, and replies, the value of dislikes couldn’t be retrieved. According to YouTube policies of 2021, the count of dislikes will not be public anymore.

### 3.3 Data Preprocessing

Since the raw data contains duplicate new data, all duplicate news stories are first removed. The elimination process is maintained as such so that the data collected later remains since it may reach more views than the previously collected one. The distinct number of news in the dataset is 16,016.

The raw data is stored in a JSON file. The date was stored as a string which was later converted as a Date type using the Pandas `to_datetime()` method. Videos on YouTube that receive more than 1,000 views and 100,000 views are sometimes annotated with the Greek letters K and M, respectively. A similar concept applies to video likes as well. To convert the object type data to numeric form, the Greek annotations are replaced by numbers. After that, the data is sorted based on their publishing date in ascending order.

While collecting the data, all the comments, replies, and respective information were stored as a single attribute. At this point, each comment and reply are separated along with the author’s information, time, and likes. To assure the privacy of the author, their names are replaced with annotated values for publicly available datasets. Data is then stored in an Excel file for further implementation.

The dataset is a Bengali dataset which is the novelty of this work. However, an English-translated version is also published for research purposes. To translate the Bengali titles, descriptions, comments, and replies, Google Sheets are used as a tool. From Google Sheets, we used the formula of “GOOGLETRANSLATE” which takes three parameters to translate a given context. The first one is the text that we want to translate, the second parameter is the source language for the text, and the third parameter is the target language. For the proposed dataset, the source language was Bengali (bn) and the target language was English (en).

For publication and censorship issues, we have replaced all the commentator’s information with a unique number so that no one can retrieve it. After completing the basic processing, the

scope of the dataset is compared to some other relevant available datasets and features in Table 3.2.

## 3.4 Data Annotation

The raw data that we collected was not labeled with the corresponding sentiment. Since the size of the dataset is quite large, we preferred automated sentiment labeling over manual annotation. For labeling the data, we tried several machine learning and language models to identify the best-suited one. In this chapter, we are going to implement different models to annotate our dataset for sentiment modeling. Then, we prepared a comparative analysis to identify the model that performs best for annotation.

### 3.4.1 Exemplary Dataset

To check the model performance, we need to use an annotated dataset. A sample-labeled dataset has been used to test the models. The SentNoB dataset contains public comments in Bengali from social media from 13 different domains that include politics, education, agriculture, and much more [53]. There are three polarity labels in the dataset where 0 refers to neutral sentiment, 1 refers to positive sentiment, and 2 is for negative. Thus, the dataset is tripolar.

The dataset contains 12,575 samples for training, 1,567 for validation, and 1,586 samples for testing. Out of the total 15,728 samples in the data, 3,609 samples were labeled as neutral, 6,410 samples were positive, and 5,709 samples were negative. The ratio of the labels identifies the dataset as an imbalanced one.

To balance the dataset “RandomOverSampler” was used from “imblearn” package. After sampling the data, the total number of samples becomes 19,230 with 6,410 samples for each class. We have combined all of them and used 80% of the data for training purposes and the rest 20% of the data for testing.

### 3.4.2 Model Selection

To start with, we have implemented 7 machine learning models to determine the performance. The models we used are Multinomial Naive Bayes, Logistic Regression, Decision Tree, Random Forest, KNN, and Support Vector Classifier (both for linear and RBF kernel). First, all the unnecessary punctuations were removed using the natural language toolkit (nltk) regular expression. The Python package for “bangla-python” was used to remove the stopwords [54]. Using the BanglaStemmer package, the words are converted to their root word. These steps were used as preprocessing to reduce the complexity of the data. For feature extraction, a TF-IDF

**Table 3.3:** Model accuracy result

| Models                             | 1-gram | 2-gram       | 3-gram |
|------------------------------------|--------|--------------|--------|
| Multinomial Naive Bayes            | 79.62  | <b>83.52</b> | 82.32  |
| Logistic Regression                | 76.0   | 81.72        | 82.87  |
| Decision Tree                      | 77.15  | 77.56        | 77.93  |
| Random Forest                      | 82.76  | 82.71        | 83.02  |
| KNN                                | 62.4   | 63.26        | 63.31  |
| Support Vector Classifier (Linear) | 68.56  | 69.99        | 71.35  |
| Support Vector Classifier (RBF)    | 70.88  | 74.75        | 76.44  |
| BERT                               | 45.8   |              |        |
| BanglaBERT                         | 68.99  |              |        |
| Azure Sentiment Analysis           | 45.78  |              |        |

vectorizer was used. To consider the continuous sequence of words, we used n-gram models for unigrams, bigrams, and trigrams. The size of the feature vector for 1-gram, 2-gram, and 3-gram methods are 27,530, 151,721, and 307,317 respectively. The accuracy of the machine learning models is represented in Table 3.3. Traditional machine learning models have been implemented here due to the high dimension of the data.

After implementing the machine learning models, two language models were also used. We have used Bidirectional Encoder Representations from Transformers (BERT) and BanglaBERT. We have used the BERT base from the BERT language model. The output sentiment scored from 1 to 5 which we later segmented as negative, neutral, and positive. A similar approach is applied to the BanglaBERT<sup>1</sup> model. The accuracy for BERT and BanglaBERT language models are 45.8% and 68.99% respectively.

There is another sentiment analysis tool developed by Microsoft named Azure Sentiment Analysis. The tool offers remarkable accuracy for the English text dataset [55, 56]. However, for our Bengali reference dataset, the accuracy of the tool was 45.78%.

Among all the machine learning and large language models (LLMs), Multinomial Naive Bayes (MNB) outperformed the rest for 2-gram methods with an accuracy of 83.52%. Thus, MNB was used to annotate our proposed dataset. We have used annotation for three features in the data: title, comment, and reply corresponding to each news.

<sup>1</sup><https://huggingface.co/ka05ar/bangla-bert-base-VITD>

## 3.5 Summary

The comparison demonstrates that the proposed one outperformed the others in terms of data volume, categories, and attributes. A corresponding article is published proposing our dataset in **Elsevier Data in Brief** journal [57]. For model selection, considering all the results from different models and tools, we noted the maximum accuracy using the Multinomial Naive Bayes for the Bigram method. Therefore, we decided to implement the model to annotate our dataset.

## Chapter 4

# Temporal Analysis

This chapter will describe some of our analysis of the dataset we prepared. Later, it focuses on the time-based analysis of the data. In temporal analysis, the pattern of the data respective to time has been incorporated. Here, we tried to depict how time and different events have affected shifts in the reaction of people to news topics. To conduct temporal analysis, both year-based and topic-based analysis have been used. Another segment is included to describe the analytics of the data other than year and topic-based analysis. The sentiment of regular and irregular commentators has also been discussed.

### 4.1 Data Analysis

Data analysis has a significant role in generating intuition about the data. Before moving on to temporal analysis, we incorporated some analysis of data through statistical analysis and graphical representation.

#### 4.1.1 Statistical Analysis

Our dataset contains various news articles data and responses from 21 June 2017 to 23 May 2023. We prepared a bar plot in Figure 4.1 to generate the frequency of comments in news data for each year. According to the diagram, the number of responses in news keeps increasing as the year progresses. The year 2017 has the minimum number of comments whereas the year 2023 has the maximum. The number of comments shows that viewers are becoming more engaged with YouTube news reports.

The ratio of views and comments was also calculated to understand how they are related to each other. A corresponding diagram is presented in Figure 4.2 which also depicts an increasing pattern in view-comment ratio. Some possible assumptions for the change could be that the

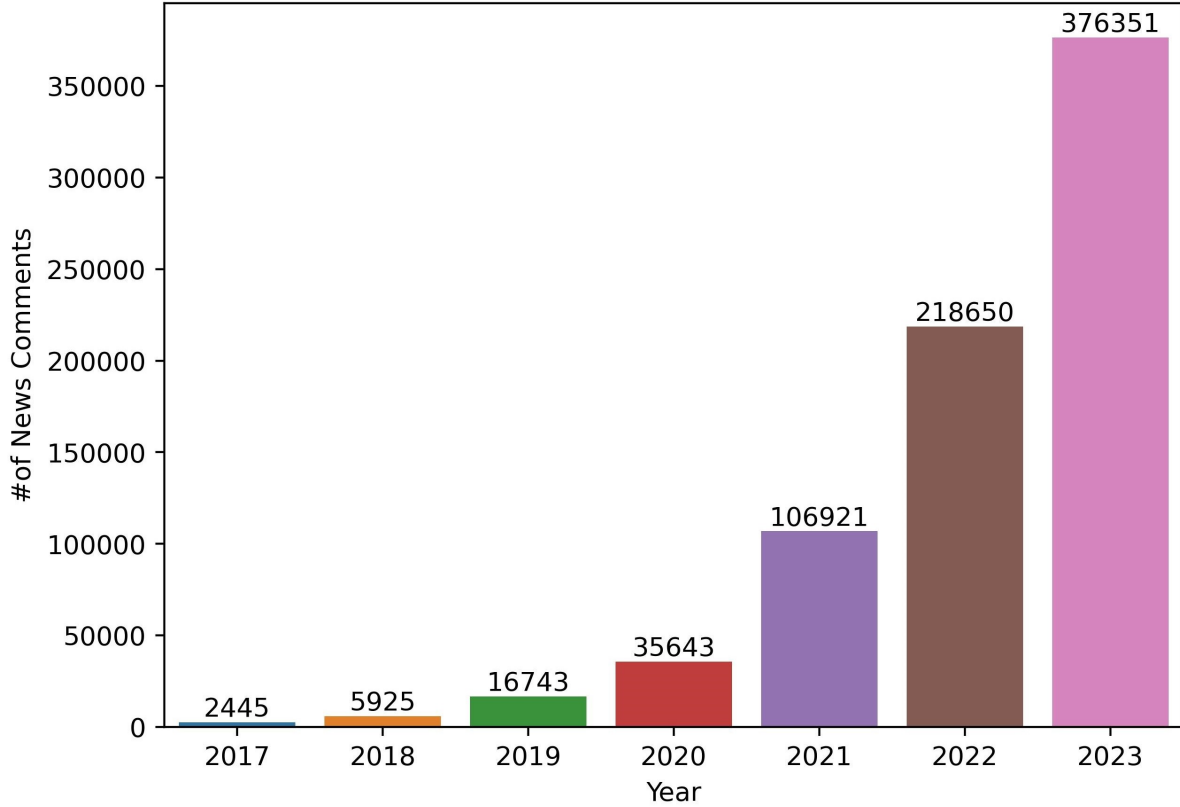


**Figure 4.1:** Number of news comments in the dataset from 2017 - 2023

number of users has increased over time, people have become more interactive on YouTube, people don't hesitate to express their opinions on public platforms, and many more.

According to Figure 4.1, the year 2023 has the maximum number of samples in the data with 3,76,351 comments. To identify the time frame where the highest news was published, a percentage is generated in Table 4.1. The data denotes that 30.7% of the news was published in March 2023. From January to March, the rate of publishing news was increasing. Whereas, the percentage declined during April.

Figure 4.3 displays the results of a correlation analysis that was carried out to determine the pattern of the data's numerical features. The correlation value depicts that all the attributes have a positive correlation among them. The maximum correlation is between the view of the videos and likes which is 0.63.

As discussed earlier, we have collected the news from 19 different categories for keyword-based searching. A frequency distribution is shown in Figure 4.4. From the diagram, we can see that the maximum number of news are from the categories **Politics** and **Bangladesh** whereas the categories **Crime**, **Opinion**, **Lifestyle**, and **Special News** have comparatively lower numbers.

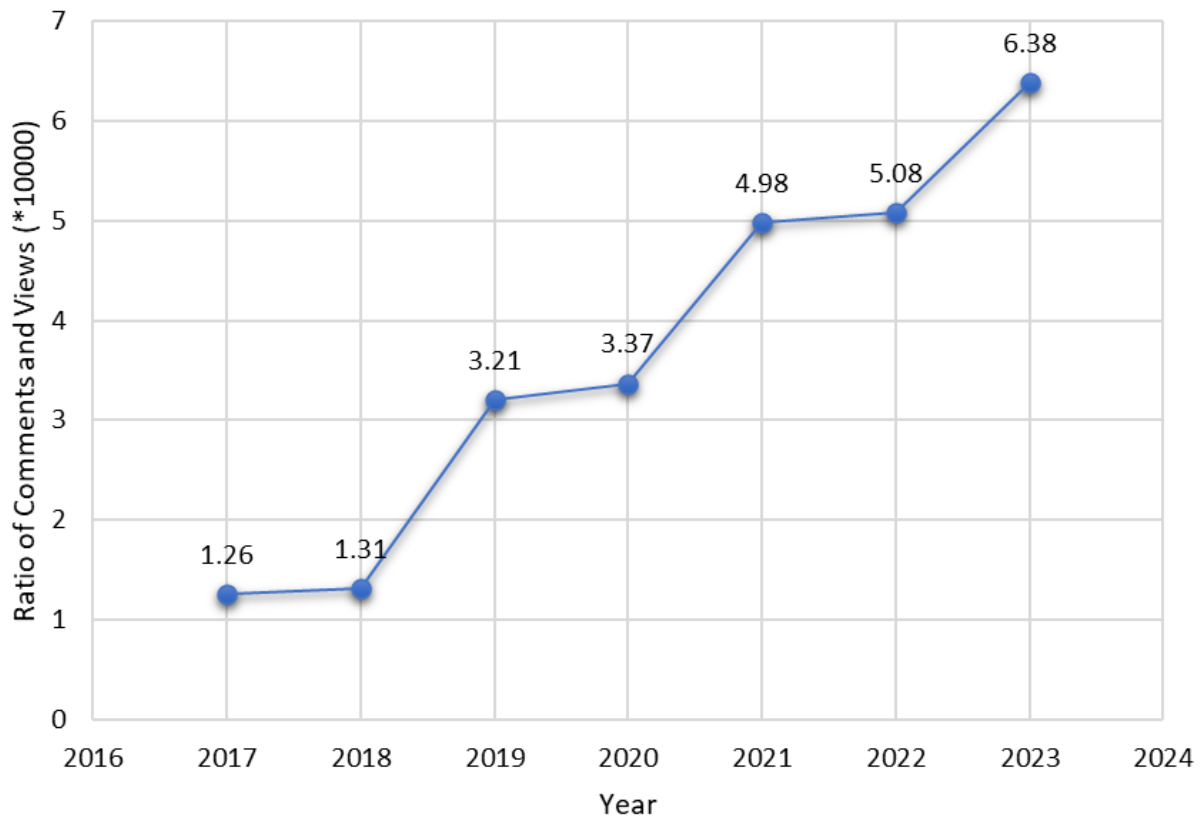


**Figure 4.2:** The ratio of comments and views for the data

of news data.

#### 4.1.2 Descriptive Analysis

In Figure 4.5, a word cloud was plotted on the hashtags used in the data to identify the most frequently used words. In some news, people don't prefer showing their opinions. There are a total of 1,453 news articles found with no comments.

Detailed information for category-based news is depicted in Table 4.2. It highlights the total number of comments for each category and the maximum number of comments respectively. The table also demonstrates the related keywords corresponding to the news that gained maximum comments.

We have identified an interesting intuition about the people who comment frequently. The number of unique authors in the dataset is 242,221 out of 761,627 comments. Thus, 68.20% of people comment frequently on news articles discarding no-comment data where the no-comment data is 1,051.

Table 4.1: Rate of news published in 2023

| Month'23 | Percentage of total news |
|----------|--------------------------|
| January  | 25.6%                    |
| February | 28.0%                    |
| March    | 30.7%                    |
| April    | 12.5%                    |
| May      | 3.3%                     |

### 4.1.3 Hypothesis Testing

We started with a null hypothesis ( $H_0$ ) which is the number of comments doesn't increase as the views for a particular video increases. Thus, the alternate hypothesis ( $H_1$ ) is the number of comments increases as the views for a particular video increase. After performing the test using Pearson Correlation, we got a statistical result of 0.60 with a p-value of 0. Since the p-value is less than the significance level of 95%, we reject the null hypothesis. Thus, the number of comments depends on the views for corresponding new data.

For the next hypothesis, we started with the assumption that the number of likes doesn't increase as the views for a particular video increase ( $H_0$ ). The result of the test is 0.50 with a p-value of 0. Thus, we reject the null hypothesis and conclude that the number of likes is positively related to the views for a particular video.

## 4.2 Year-oriented Analysis

The dataset represents data within the time frame of 2017 to 2023. The news and comments over time are not constant. Figure 4.6a and 4.6b depict how the number of news and comments changed from 2017 to 2023 respectively. From the diagram, we can see that public engagement in news videos tremendously increased in late 2020 and was at its peak in 2023. There is an unusual pattern observed in September 2018.

For a similar number of news compared to the collateral months, the number of comments increased around 10 times. After observing the news source, we figured out that the maximum number of comments came from two news articles. One is about including a Boeing 787 to Bi-man Bangladesh Airlines<sup>1</sup> (1010 comments) and another is a political movement of Bangladesh Nationalist Party, BNP (major opponent political party of Bangladesh)<sup>2</sup> (873 comments) following the national election of Bangladesh which was held on 30 December 2018 [58]. A similar

---

<sup>1</sup>Boeing 787 Dreamliner | আকাশবীণার ককপিটে শেখ হাসিনা | Sheikh Hasina

<sup>2</sup>Exclusive: জোগানে মুখর বিএনপির জনসভাস্থল | ৫ দফা দাবি তোলার কথা | BNP Assembly | Somoy TV

## 4.2 Year-oriented Analysis

**Table 4.2:** Category-wise news information

| Category      | Number of News | Total Comments | Maximum Number of Comments | News Topic for Maximum Comments                    |
|---------------|----------------|----------------|----------------------------|----------------------------------------------------|
| Politics      | 760            | 26,158         | 1,258                      | Sohel Taj, Politician, Fitness And Health Activist |
| Bangladesh    | 760            | 12,441         | 1,186                      | France News, Pinaki Bhattacharya                   |
| World         | 655            | 14,758         | 1,192                      | Afghanistan News, World Bank, International News   |
| Crime         | 191            | 33,323         | 1,239                      | Kishor Gang, Khulna Court, Khulna News             |
| Trade         | 613            | 13,005         | 1,063                      | Export Goods Theft, Dhaka News                     |
| Opinion       | 144            | 25,700         | 1,945                      | Kurigram News                                      |
| Sports        | 659            | 33,226         | 1,123                      | Eleta Kingsley, Sports News                        |
| Entertainment | 394            | 28,691         | 1,729                      | BTS, Interview on BTS                              |
| Job           | 588            | 13,882         | 1,339                      | Chanchal Chowdhury                                 |
| Lifestyle     | 171            | 8,042          | 1,423                      | Army Life, Bangladesh Army                         |
| Education     | 541            | 41,589         | 1,103                      | Dipu Moni, Education Minister                      |
| Election      | 652            | 31,125         | 1,275                      | DSCC Election                                      |
| FIFA          | 334            | 28,828         | 899                        | Argentina vs Saudi Arabia                          |
| Earthquake    | 314            | 14,797         | 1,236                      | Poran, Bangladesh Police, New Market               |
| Price Hike    | 443            | 43,882         | 1,355                      | Bongo Bazar Fire, Dhaka Fire News                  |
| Padma Bridge  | 308            | 39,176         | 1,103                      | Padma Bridge Visit                                 |
| Metro Rail    | 312            | 15,687         | 790                        | Metro Rail, Public Reaction                        |
| Latest News   | 517            | 27,834         | 1,180                      | Alcor, Life Extension Foundation                   |
| Special News  | 93             | 7,881          | 1,173                      | Ramadan Special Moon                               |

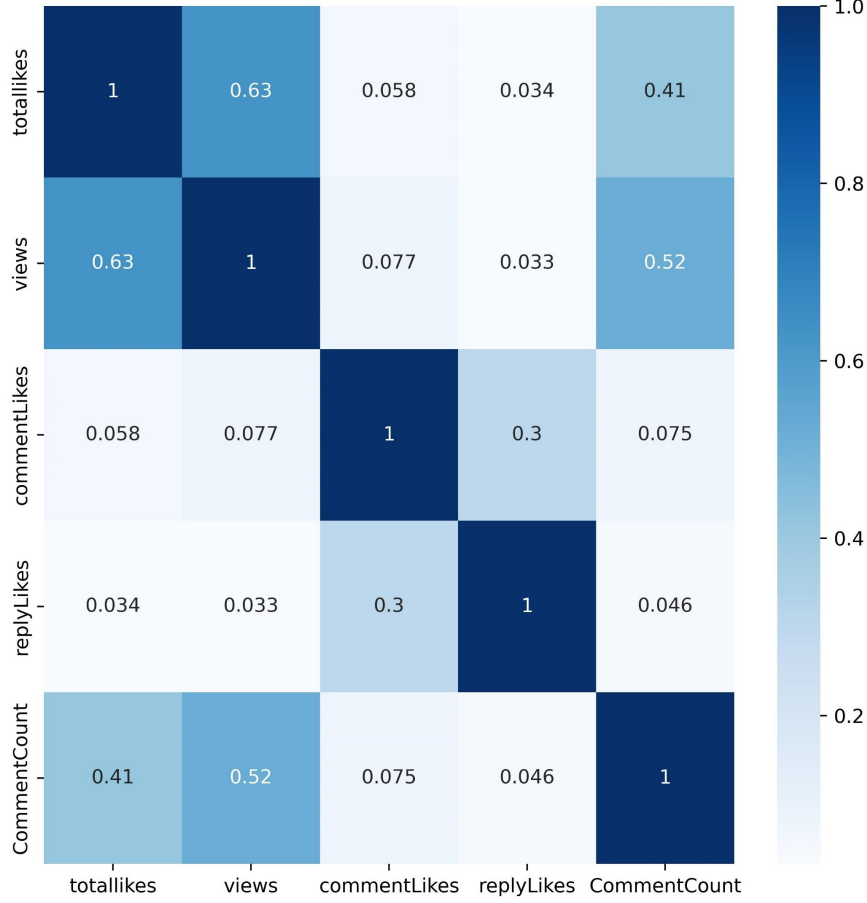


**Figure 4.3:** Correlation between the numerical attributes

pattern is also seen from March 2019 to May 2019. Public reaction increased for a lower number of news. Here, the maximum comments came from a speech by Nurul Haq Nur (26<sup>th</sup> Vice President of Dhaka University Central Students' Union - DUCSU) for the honorable Prime Minister of Bangladesh<sup>1</sup> (1442 comments) and a crime report which is about a rape incident in March 2019 from Feni, Bangladesh<sup>2</sup> (915 comments).

In December 2020, both the news data and public engagement were raised. Maximum public comments are noted from two news articles regarding Padma Bridge where the construction development of the bridge<sup>3</sup> (944 comments) and the toll payment<sup>4</sup> procedure (790 comments) are presented respectively. Padma Multipurpose Bridge (Padma Bridge) is one the most important

<sup>1</sup>প্রধানমন্ত্রী শেখ হাসিনার মাঝে আমার মায়ের প্রতিচ্ছবি খুঁজে পাই | nurul haq nur

<sup>2</sup>ফেনীর নুসরাত হত্যা ঘটনার লোমহর্ষক বর্ণনা দিলো পিবিআই | nusrat jahan rafi | omar faroque

<sup>3</sup>সব কটাক্ষের জবাব দিয়ে দৃশ্যমান পদ্মা সেতু! | Padma Bridge | Khaleda Zia | Somoy TV

<sup>4</sup>পদ্মা সেতুতে টোল দিতে থামতে হবেনা কোন গাড়িকে | Padma Bridge



## 4.2 Year-oriented Analysis

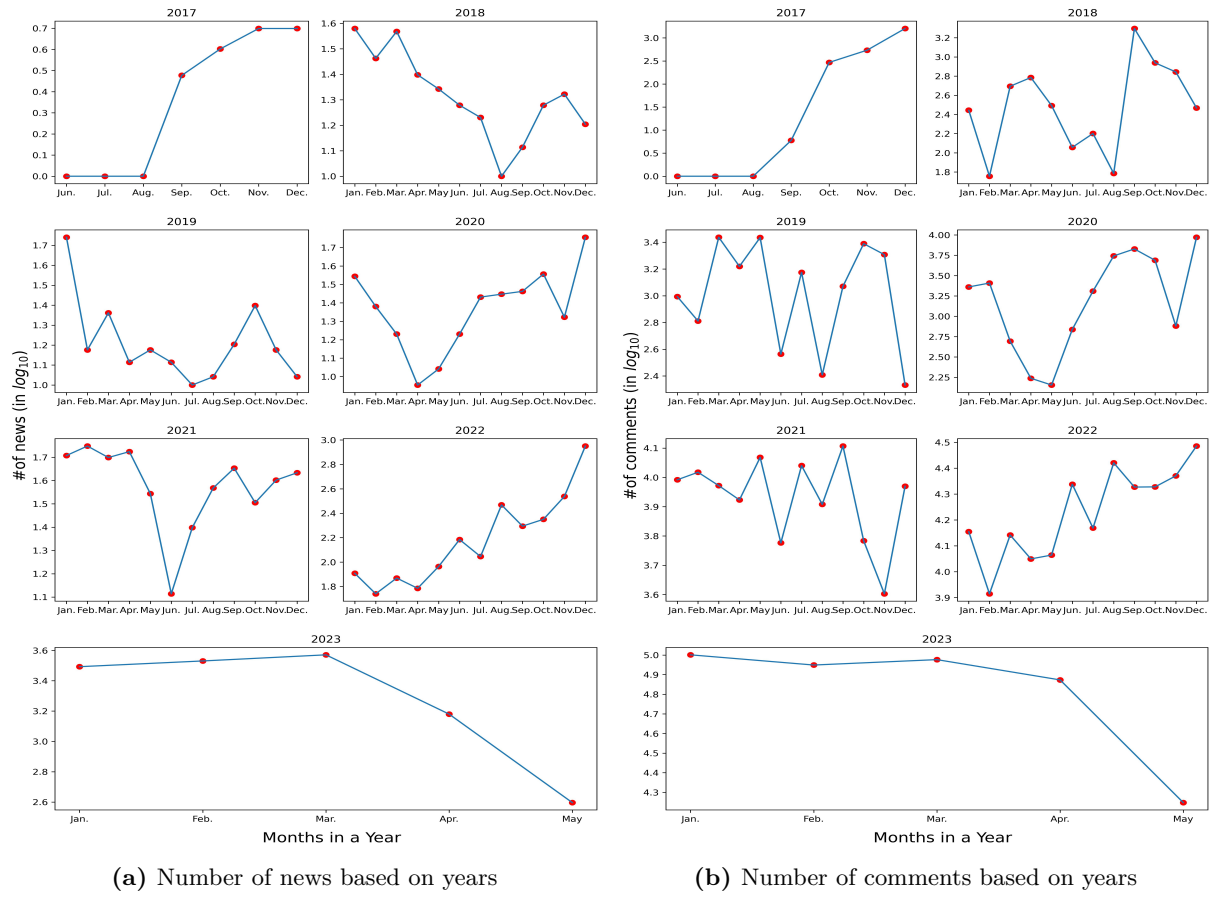


**Figure 4.6:** Year-wise news and comments information

and challenging construction projects of Bangladesh with a two-level road-rail across the Padma River constructed by China Major Bridge Engineering Company Ltd. It is the longest bridge in Bangladesh consisting of 41 sections with a total length of 6.15 km. It is the longest bridge in Bangladesh. The construction was challenging due to the width and depth of the Padma River. Padma Bridge has the world's deepest pile of any bridge at 127 meters. It connects the southwest of the country to the northern and eastern regions. Though the construction of the bridge was expected to commence by early 2011 and be ready for major completion by 2013, the bridge was inaugurated on 25 June 2022. Following the bridge's opening, economic activity in the southwestern regions is expected to be boosted. The bridge is expected to boost the GDP of Bangladesh by 1.23% [59].

While observing the year 2021 the number of comments in November showed a lower pattern compared to other months for a similar number of news data. It means people became uninterested in showing their reaction to this news. The news topics during this period are mostly regarding the Union Parishad (UP) Election, Dhaka Metro Rail, and Padma Bridge where the least public reaction is noted from the news regarding UP Election<sup>1</sup> (1 comment). The pattern of comments for the years 2022 and 2023 is almost similar to the number of news data. Like Padma Bridge, Dhaka Metro Rail is another notable construction project in Bangladesh to ensure a mass rapid transit system in Dhaka. With the goal of Dhaka's 20-year-long Strategic Transport Plan, the Japan International Cooperation Agency was invited by the government of Bangladesh to survey the transport system of Dhaka in 2009–2010. In 2012, the project was approved with a loan agreement between the Bangladeshi government and the Japanese International Cooperation Agency. The construction was combined by the engineers of the Italian-Thai Development Public Company, China's Sinohydro Corporation, and a Tokyo-based construction company. On 29 December 2022, the first phase of MRT line-6 was opened for public use within the terminal Uttara to Agargaon [60].

In the dataset, the maximum number of comments appeared on the news where people gathered to see a grave with some miraculous feature. The news got a total of 1945 comments<sup>2</sup>.

### 4.3 Topic-oriented Analysis

To observe how the sentiment of published news and comments change over time, we have done some temporal analysis for specific topics. We tried to plot the sentiment for the news related to the Padma Bridge from 2017 to 2023 in Figure 4.7.

---

<sup>1</sup>তৃতীয় ধাপে এক হাজার ইউনিয়ন পরিষদ নির্বাচনের ভোটগ্রহণ শুরু! | up election | 2021 | somoy tv

<sup>2</sup>কুড়িগ্রামে কবর দেখতে উৎসুক জনতার ভিড় | Kurigram News | Somoy TV

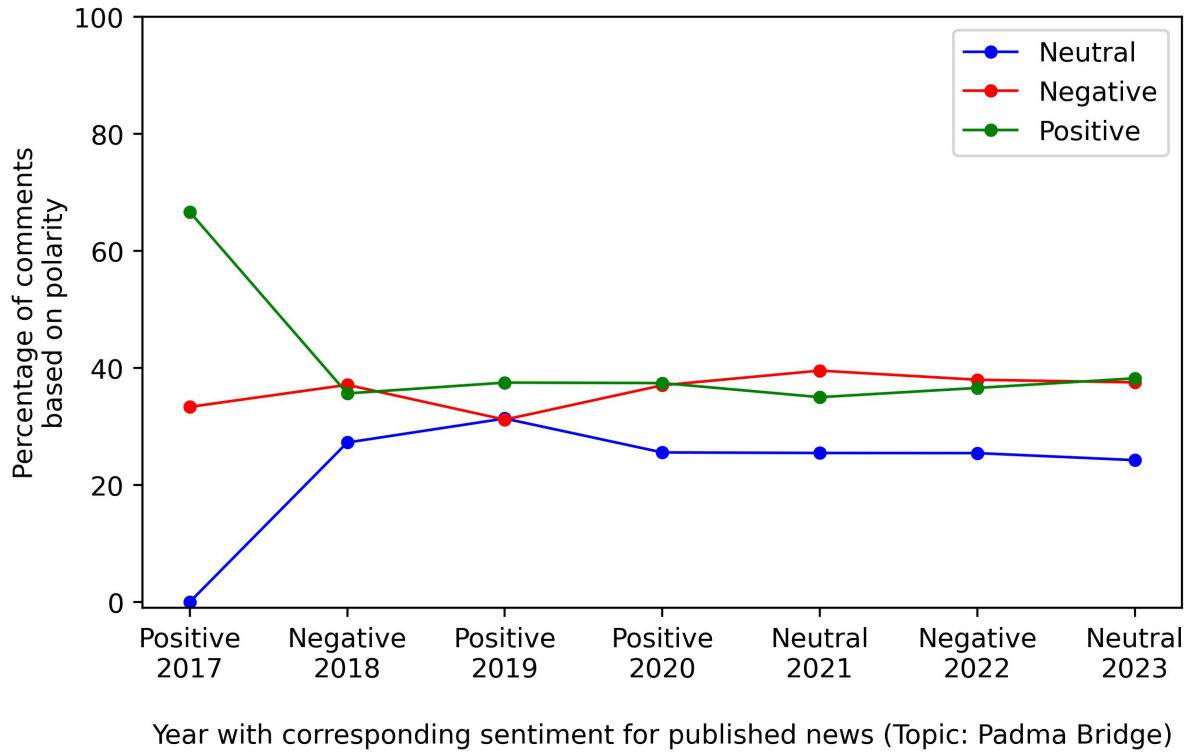


**Figure 4.7:** Sentiment on Padma bridge from 2017 - 2023

The news sentiment varied over time. It started with a highly positive sentiment in 2017. The corresponding news was about the development of Padma Bridge<sup>1</sup> and the architectural design of Padma Bridge<sup>2</sup>. However, the positive sentiment dropped between the year 2018 to 2023. During 2021 the negative (1,954 comments) sentiment became higher than the positive one (1,730 comments). From the percentage of comments polarity, we can notice that in 2018, positive public comments slightly decreased, and negative comments increased. One possible reason for this change could be the completion of the works of Padma Bridge which was supposed to be completed by 2018 and news published regarding the construction complexity of the bridge<sup>3</sup>. Padma Bridge was opened for public use on June 25, 2022, and the positive sentiment started to develop after that.

Figure 4.8 represents the corresponding news and comments sentiment for the news on Bangladesh. For the news sentiment, it started with a negative polarity in 2018. The comments became positive in 2019 when the news was about acquiring the 45th position of Bangladesh's

<sup>1</sup>পদ্মা সেতুর সর্বশেষ অগ্রগতি! | Padma Bridge Latest Update | Somoy TV

<sup>2</sup>পরিবর্তন হচ্ছে পদ্মা সেতুর নকশায়! | Padma Bridge Latest News | Somoy TV

<sup>3</sup>নির্মাণে জটিলতা | উপায় খুঁজে পাচ্ছে না প্রকৌশলীরা | Padma Bridge Latest Update | Somoy TV

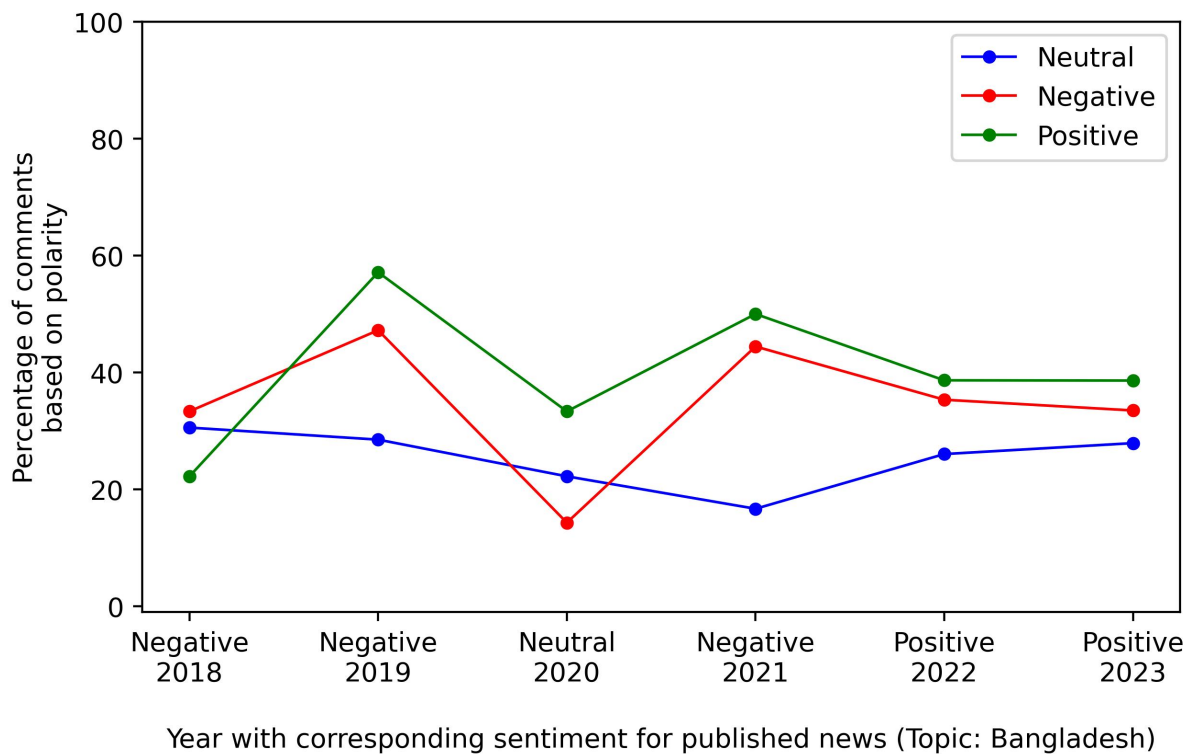


**Figure 4.8:** Sentiment on Bangladesh news from 2018 - 2023

GDP in Asia<sup>1</sup>. However, the comment's polarity was constantly positive afterward. We noticed a change in the comment sentiment in 2020 where the number of negative comments dropped. To investigate the possible reasoning, we tried to explore the news data for 2020 and figured out that most of the positive comments were sourced compared to negative news about reopening the flights after the COVID-19 period<sup>2</sup>.

Similarly, Figure 4.9a, 4.9b, 4.9c, and 4.9d represents the corresponding news and comments sentiments for the news regarding Earthquake, Education, Lifestyle, and International news respectively. In such a way, we can analyze the public reaction to a particular news topic.

## 4.4 Others

Another interesting finding from our data is how often the users return to share their opinions. After excluding the news with no comment, we identified that only 31.80% of the commentators are unique, and 68.20% people commented more than one time. Then, we tried to identify the

<sup>1</sup>এশিয়ার ৪৫টি দেশের মধ্যে বাংলাদেশ সর্বোচ্চ পর্যায়ে পৌঁছেছে' | AK Abdul Momen

<sup>2</sup>আপাদত ফ্লাইট চলাচল সম্ভব নয় | AK Abdul Momen | Somoy TV

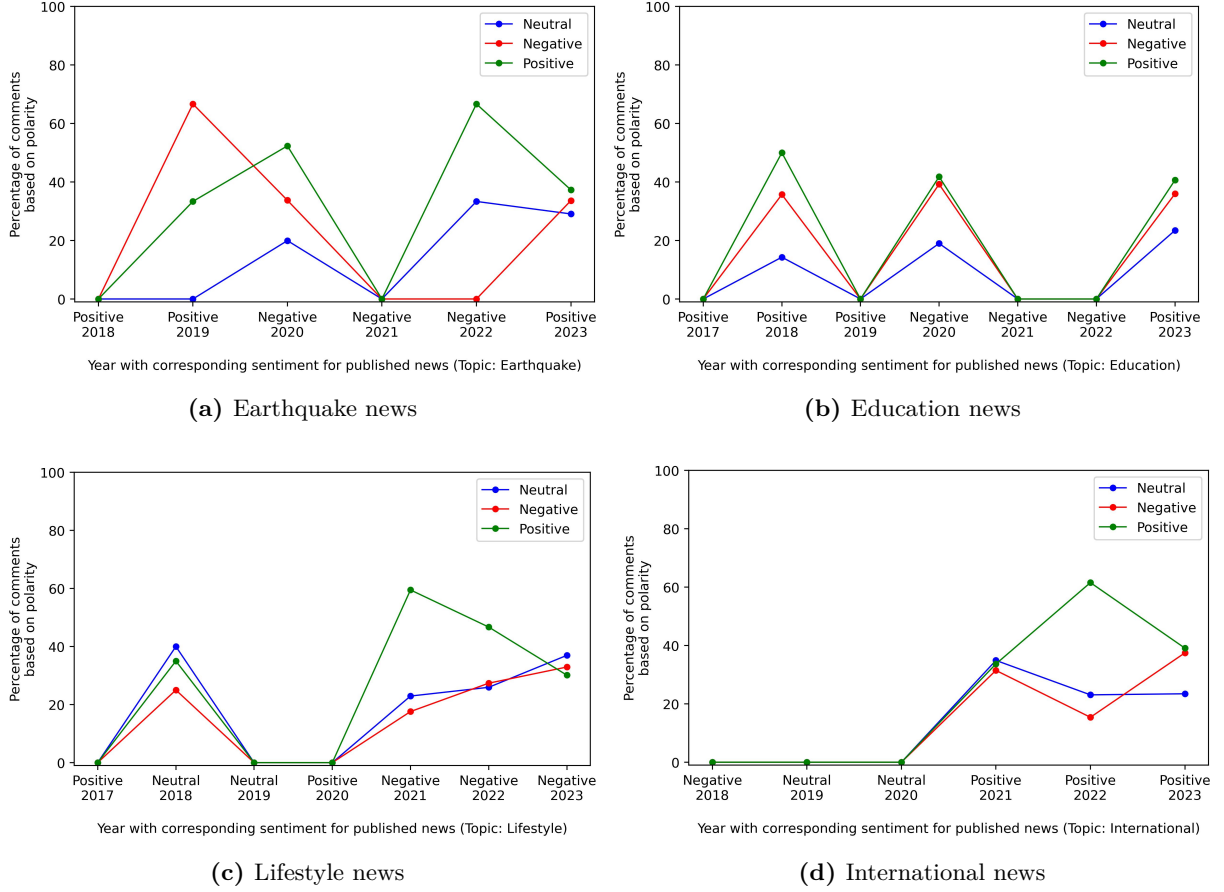


**Figure 4.9:** Temporal analysis of sentiments for corresponding news

pattern of their comments for particular news videos. In Figure 4.10, the ratio for frequent and non-frequent users is depicted for the news categorized as neutral, positive, and negative news data. The ratio narrates that the ratio between regular and irregular users for all sentiment of news data is almost identical.

Figure 4.11 represents the comments received for the news data on each polarity. The ratio shows that the maximum number of news data collected have a negative sentiment so as the comments.

## 4.5 Result Analysis

In this study, we implemented 7 machine-learning models and 3 language models for sentiment analysis on a labeled dataset. The N-grams method has been applied to select successive items in a text. According to the result, Multinomial Naive Bayes (MNB) for the 2-gram method

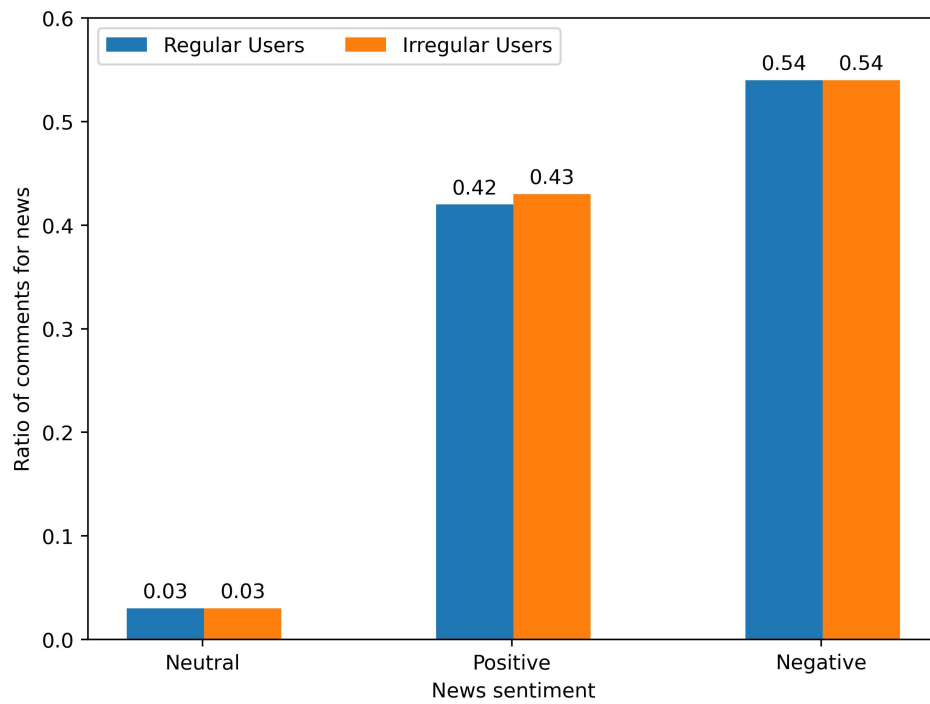


**Figure 4.10:** Ratio of comments between regular and irregular user

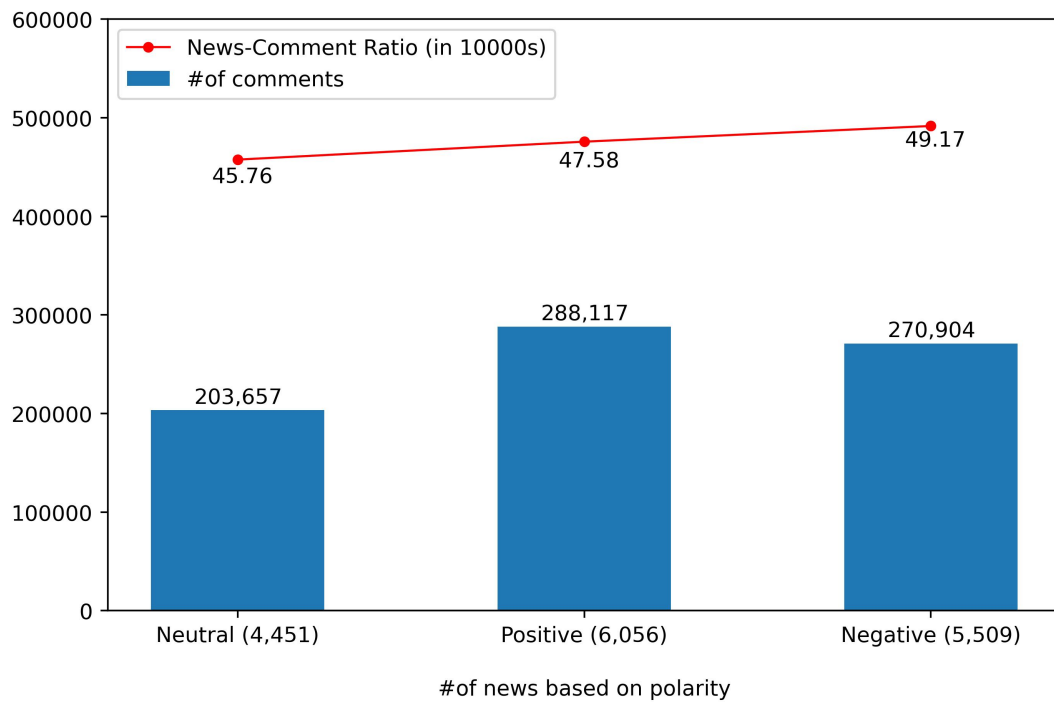


**Figure 4.11:** Sentiment-based ratio for news and comments

outperformed the rest for news data annotation, Figure 4.12. Thus, we used MNB to label the news titles, comments, and replies.

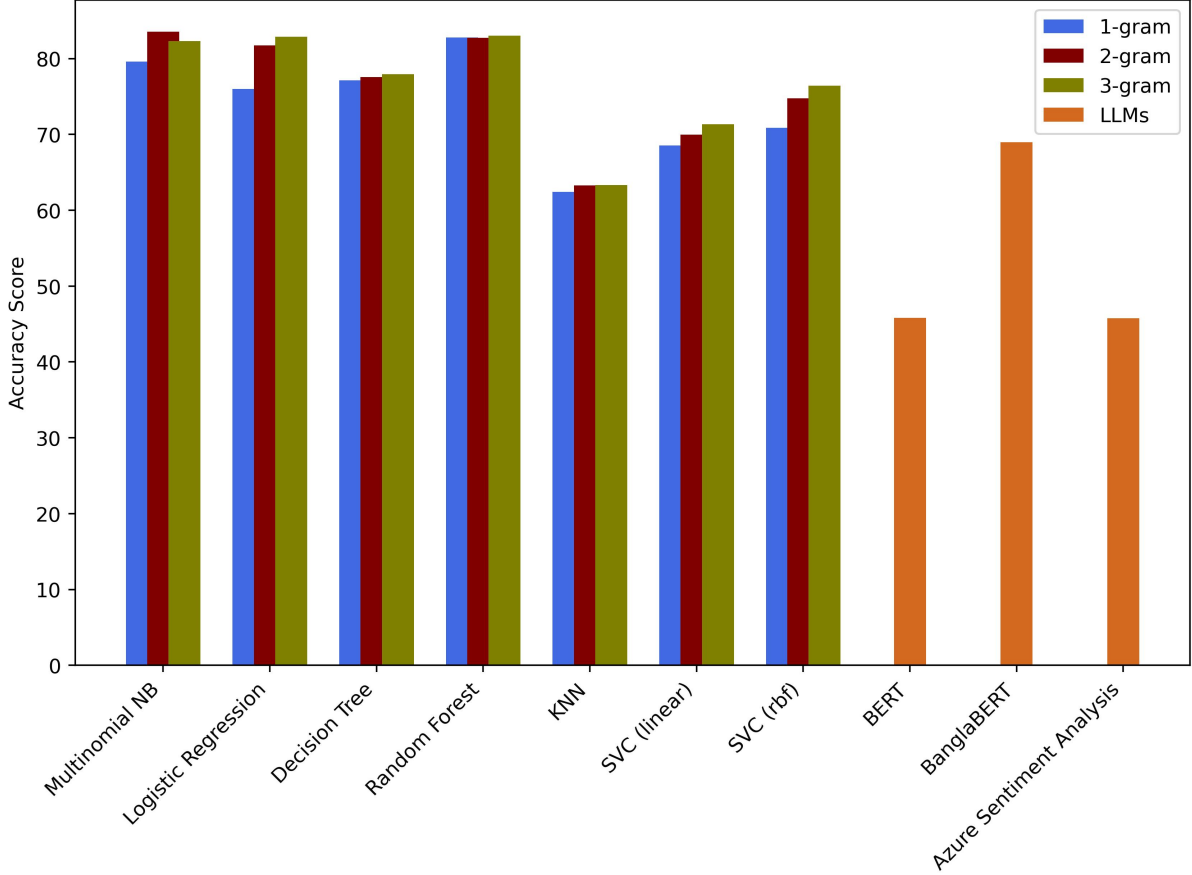


**Figure 4.12:** Result of sentiment annotation Models

After annotation, the temporal pattern of the data has been observed. The temporal analysis of the data for different topics depicts that public sentiment varies as time goes on. There are events that trigger the corresponding polarity to shift from positive to negative and vice-versa. Documenting and understanding these events is important for contextualizing sentiment shifts or specifying if certain types of events consistently lead to positive or negative sentiment. For example, we have seen that initially, the sentiment related to the Padma Bridge was more negative. However, people showed a positive reaction as the development of the bridge continued. Such intuition may help the government and policymakers with early prediction of public reaction to their action in the development sector.

According to the data, people are more likely to comment on negative news compared to positive and neutral. There is a repetitive pattern found among the commentators. The pattern

shows that most of the users come back to share their opinions on the portal. These frequent users mostly share their opinions on negative news data. The pattern of temporal analysis also reveals that the news and comments sentiments are mostly bipolar. Thus, the number of neutral sentiments compared to positive and negative sentiments is negligible.

## 4.6 Summary

Through descriptive and statistical analysis, we now have some idea about our data and its properties. This intuition may later guide us to focus on significant features. The temporal analysis represents that public sentiment changes over time based on some triggering event or topic. However, people tend to show positive sentiment more compared to negative and neutral sentiment in the news videos. Some users frequently comment on the news and are more likely to comment negatively.

## Chapter 5

# Conclusion and Future Work

### 5.1 Conclusion

To conclude, the analysis successfully linked specific occurrences to changes in sentiment, highlighting the interconnectedness between real-world events and public perception.

The study proposes a dataset containing public comments on published Bengali news from 2017 to 2023. The dataset contains a total of 8 features for each news sample. For sentiment polarity annotation, after implementing different machine learning and large language models (LLMs) and comparing the results, an optimal model for sentiment analysis for news data has been proposed.

The study exhibits notable temporal patterns in public sentiment over the specified time frame. Two topics have been focused on discussing the temporal pattern of the data (Padma Bridhe and Bangladesh). Events were identified as significant triggers for shifts in public sentiment for the news regarding these topics. Another four topic-based sentiment temporal pattern is depicted.

Visualizations, such as time series plots and heatmaps, were utilized to effectively communicate temporal patterns and the impact of events on sentiment. These visual aids enhance interpretability and facilitate a clearer understanding of the results. For any transition point of the polarity, we can further investigate the news topic to answer the reason for sentiment change.

In summary, the analysis not only identified an effective model for sentiment analysis but also uncovered temporal patterns and event-driven sentiment changes in the news dataset. These findings contribute to a deeper understanding to provide valuable insights for future research, decision-making, and applications in sentiment analysis for a time range.

## 5.2 Future Work

The incorporation of temporal patterns and event-driven sentiment changes adds depth to the understanding of public sentiment dynamics over the specified time frame. Having a better intuition regarding the fluctuations is required for gaining insights into how sentiment evolves over time. Specifying the nature of these patterns - whether sentiment changes were gradual, sudden, or periodic, can lead to further areas of research.

In the future, we plan to create a recommendation model based on the dataset to predict whether a particular user will react to a specific news or not. Predicting the plausible reaction to news data is also on our list.

# Bibliography

- [1] Wikipedia contributors, “Language processing in the brain — Wikipedia, the free encyclopedia.” [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Language\\_processing\\_in\\_the\\_brain&oldid=1136990052](https://en.wikipedia.org/w/index.php?title=Language_processing_in_the_brain&oldid=1136990052) 1
- [2] J. Hutchins, “The history of machine translation in a nutshell,” *Retrieved December*, vol. 20, no. 2009, pp. 1–1, 2005. 1
- [3] A. V. Aho, R. Sethi, and J. D. Ullman, *Compilers: principles, techniques, and tools*. Addison-wesley Reading, 2007, vol. 2. 1
- [4] N. Chomsky, “On certain formal properties of grammars,” *Information and control*, vol. 2, no. 2, pp. 137–167, 1959. 1
- [5] C. Mellish and W. Clocksin, “Programming in prolog: Using the iso standard,” 2003. 1
- [6] G. Moore, “Moore’s law made real by intel innovation,” *Intel, Santa Clara, CA*, 1965. 1
- [7] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014. 1
- [8] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “Natural language processing: an introduction,” *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011. 1
- [9] Wikipedia contributors, “Natural language processing — Wikipedia, the free encyclopedia,” 2023. [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Natural\\_language\\_processing&oldid=1141420081](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&oldid=1141420081) 2
- [10] K. Rekanar, I. R. O’Keeffe, S. Buckley, M. Abbas, S. Beecham, M. Chochlov, B. Fitzgerald, L. Glynn, K. Johnson, J. Laffey *et al.*, “Sentiment analysis of user feedback on the hse’s covid-19 contact tracing app,” *Irish Journal of Medical Science (1971-)*, pp. 1–10, 1971. 2

- [11] L. Kanal, "Patterns in pattern recognition: 1968-1974," *IEEE Transactions on information theory*, vol. 20, no. 6, pp. 697–722, 1974. [2](#)
- [12] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Knowledge-based approaches to concept-level sentiment analysis," *IEEE intelligent systems*, vol. 28, no. 2, pp. 12–14, 2013. [2](#)
- [13] J. Vizcarra, K. Kozaki, M. Torres Ruiz, and R. Quintero, "Knowledge-based sentiment analysis and visualization on social networks," *New Generation Computing*, vol. 39, pp. 199–229, 2021. [2](#)
- [14] A. Karn, A. Shrestha, A. Pudasaini, B. Mahara, and A. Jaiswal, "Statistic based sentiment analysis of social media data," *Int. Res. J. Innov. Eng. Technol*, vol. 2, pp. 28–32, 2018. [2](#)
- [15] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to sentiment analysis," in *2016 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2016, pp. 4950–4957. [2](#)
- [16] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5876–5880. [2](#)
- [17] A. Alrehili and K. Albalawi, "Sentiment analysis of customer reviews using ensemble method," in *2019 International conference on computer and information sciences (ICCIS)*. IEEE, 2019, pp. 1–6. [2](#)
- [18] F. Poecze, C. Ebster, and C. Strauss, "Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts," *Procedia computer science*, vol. 130, pp. 660–666, 2018. [2](#)
- [19] L. Abualigah, H. E. Alfar, M. Shehab, and A. M. A. Hussein, "Sentiment analysis in healthcare: a brief review," *Recent advances in NLP: the case of Arabic language*, pp. 129–141, 2020. [2](#)
- [20] M. T. Hannan and N. B. Tuma, "Methods for temporal analysis," *Annual review of sociology*, vol. 5, no. 1, pp. 303–328, 1979. [3](#)
- [21] J. Mohammed and A. O. Sulyman, "Spatio-temporal analysis of bida housing market using geographic information system," 2019. [4](#)

- [22] A. B. Silva, J. Arigony-Neto, M. H. Braun, J. M. A. Espinoza, J. Costi, and R. Jaña, “Spatial and temporal analysis of changes in the glaciers of the antarctic peninsula,” *Global and Planetary Change*, vol. 184, p. 103079, 2020. [4](#)
- [23] L. Soltanisehat, R. Alizadeh, H. Hao, and K.-K. R. Choo, “Technical, temporal, and spatial research challenges and opportunities in blockchain-based healthcare: A systematic literature review,” *IEEE Transactions on Engineering Management*, vol. 70, no. 1, pp. 353–368, 2020. [4](#)
- [24] M. Yang, Y. Ren, and G. Adomavicius, “Understanding user-generated content and customer engagement on facebook business pages,” *Information Systems Research*, vol. 30, no. 3, pp. 839–855, 2019. [5](#)
- [25] P. Chauhan, N. Sharma, and G. Sikka, “The emergence of social media data and sentiment analysis in election prediction,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 2601–2627, 2021. [5](#)
- [26] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, “A survey on concept-level sentiment analysis techniques of textual data,” in *2021 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2021, pp. 0285–0291.
- [27] Z. Drus and H. Khalid, “Sentiment analysis in social media and its application: Systematic literature review,” *Procedia Computer Science*, vol. 161, pp. 707–714, 2019. [5](#)
- [28] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, “Covidsentis: A large-scale benchmark twitter data set for covid-19 sentiment analysis,” *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 1003–1015, 2021. [5](#)
- [29] M. O. Lwin, J. Lu, A. Sheldenkar, P. J. Schulz, W. Shin, R. Gupta, and Y. Yang, “Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends,” *JMIR public health and surveillance*, vol. 6, no. 2, p. e19447, 2020. [5](#)
- [30] R. H. Ali, G. Pinto, E. Lawrie, and E. J. Linstead, “A large-scale sentiment analysis of tweets pertaining to the 2020 us presidential election,” *Journal of big Data*, vol. 9, no. 1, pp. 1–12, 2022. [5](#)
- [31] H. Jang, E. Rempel, I. Roe, P. Adu, G. Carenini, and N. Z. Janjua, “Tracking public attitudes toward covid-19 vaccination on tweets in canada: Using aspect-based sentiment analysis,” *Journal of Medical Internet Research*, vol. 24, no. 3, p. e35016, 2022. [5](#)

- [32] S. Žitnik, N. Blagus, and M. Bajec, “Target-level sentiment analysis for news articles,” *Knowledge-Based Systems*, vol. 249, p. 108939, 2022. [5](#)
- [33] S. Park, S. Strover, J. Choi, and M. Schnell, “Mind games: A temporal sentiment analysis of the political messages of the internet research agency on facebook and twitter,” *New Media & Society*, vol. 25, no. 3, pp. 463–484, 2023. [10](#)
- [34] T. Hu, B. She, L. Duan, H. Yue, and J. Clunis, “A systematic spatial and temporal sentiment analysis on geo-tweets,” *Ieee Access*, vol. 8, pp. 8658–8667, 2019. [10](#)
- [35] A. Chatterjee and A. Das, “Temporal sentiment analysis of the data from social media to early detection of cyberbullicide ideation of a victim by using graph-based approach and data mining tools,” *Intelligence Enabled Research: DoSIER 2019*, pp. 107–112, 2020. [11](#)
- [36] S. Gupta and R. Sandhane, “Use of sentiment analysis in social media campaign design and analysis,” *Cardiometry*, no. 22, pp. 351–363, 2022. [11](#)
- [37] S. Yu, D. Eisenman, and Z. Han, “Temporal dynamics of public emotions during the covid-19 pandemic at the epicenter of the outbreak: sentiment analysis of weibo posts from wuhan,” *Journal of medical Internet research*, vol. 23, no. 3, p. e27078, 2021. [11](#)
- [38] I. Lorentz and G. Singh, “Sentiment analysis on youtube comments to predict youtube video like proportions,” 2021. [12](#)
- [39] A. A. L. Cunha, M. C. Costa, and M. A. C. Pacheco, “Sentiment analysis of youtube video comments using deep neural networks,” in *Artificial Intelligence and Soft Computing: 18th International Conference, ICAISC 2019, Zakopane, Poland, June 16–20, 2019, Proceedings, Part I 18*. Springer, 2019, pp. 561–570. [12](#)
- [40] A. Al-Laith and M. Shahbaz, “Tracking sentiment towards news entities from arabic news on social media,” *Future Generation Computer Systems*, vol. 118, pp. 467–484, 2021. [13](#)
- [41] A. Giachanou and F. Crestani, “Tracking sentiment by time series analysis,” in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 1037–1040. [13](#)
- [42] T. Carpenter and T. Way, “Tracking sentiment analysis through twitter,” in *Proceedings of the international conference on information and knowledge engineering (IKE)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2012, p. 1. [13](#)

- [43] X. Ji, S. A. Chun, and J. Geller, “Monitoring public health concerns using twitter sentiment classifications,” in *2013 IEEE International Conference on Healthcare Informatics*. IEEE, 2013, pp. 335–344. [13](#)
- [44] M. L. Jamil, S. Pais, and J. Cordeiro, “Detection of dangerous events on social media: a critical review,” *Social Network Analysis and Mining*, vol. 12, no. 1, p. 154, 2022. [13](#)
- [45] M. A. Faruque, S. Rahman, P. Chakraborty, T. Choudhury, J.-S. Um, and T. P. Singh, “Ascertaining polarity of public opinions on bangladesh cricket using machine learning techniques,” *Spatial Information Research*, pp. 1–8, 2021. [14](#)
- [46] G. M. Insights, “Youtube users statistics 2023,” 16 Nov 2023. [Online]. Available: [https://www.globalmediainsight.com/blog/youtube-users-statistics/#:\\$\sim\\$:text=YouTube%20has%20more%20than%202.70,population%20uses%20YouTube%20every%20month](https://www.globalmediainsight.com/blog/youtube-users-statistics/#:$\sim$:text=YouTube%20has%20more%20than%202.70,population%20uses%20YouTube%20every%20month) [16](#)
- [47] YouTube, “Somoy tv.” [Online]. Available: <https://www.youtube.com/@somoynews360/featured> [16](#)
- [48] A. Hande, R. Priyadarshini, and B. R. Chakravarthi, “Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection,” in *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, 2020, pp. 54–63. [19](#)
- [49] M. Hasan, L. Islam, I. Jahan, S. M. Meem, and R. M. Rahman, “Natural language processing and sentiment analysis on bangla social media comments on russia–ukraine war using transformers,” *Vietnam Journal of Computer Science*, pp. 1–28, 2023. [19](#)
- [50] M. A. Rahman and E. Kumar Dey, “Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation,” *Data*, vol. 3, no. 2, p. 15, 2018. [19](#)
- [51] M. Ahmed Masum, S. Junayed Ahmed, A. Tasnim, and M. Saiful Islam, “Ban-absa: An aspect-based sentiment analysis dataset for bengali and its baseline evaluation,” in *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*. Springer, 2021, pp. 385–395. [19](#)
- [52] M. Data, “Bengali youtube news opinion data from youtube,” 24 Nov 2023. [Online]. Available: <https://data.mendeley.com/datasets/3c3j3bkxvn/4> [19](#)
- [53] K. I. Islam, S. Kar, M. S. Islam, and M. R. Amin, “Sentnob: A dataset for analysing sentiment on noisy bangla texts,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3265–3271. [21](#)

- [54] N. Hossain, “bangla-python,” 2022. 21
- [55] L. M. Qaisi and I. Aljarah, “A twitter sentiment analysis for cloud providers: A case study of azure vs. aws,” in *2016 7th International Conference on Computer Science and Information Technology (CSIT)*. IEEE, 2016, pp. 1–6. 22
- [56] O. Harfoushi, D. Hasan, R. Obiedat *et al.*, “Sentiment analysis algorithms through azure machine learning: Analysis and comparison,” *Modern Applied Science*, vol. 12, no. 7, p. 49, 2018. 22
- [57] L. H. Chowdhury, S. Islam, and S. Shatabda, “A bengali news and public opinion dataset from youtube,” *Data in Brief*, p. 109938, 2023. 23
- [58] Wikipedia contributors, “2018 bangladeshi general election — Wikipedia, the free encyclopedia,” 2024. [Online]. Available: [https://en.wikipedia.org/wiki/2018\\_Bangladeshi\\_general\\_election](https://en.wikipedia.org/wiki/2018_Bangladeshi_general_election) 27
- [59] W. contributors, “Padma Bridge — Wikipedia, The Free Encyclopedia,” 2024. [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Padma\\_Bridge&oldid=1215126163](https://en.wikipedia.org/w/index.php?title=Padma_Bridge&oldid=1215126163) 32
- [60] Wikipedia contributors, “Dhaka metro rail — Wikipedia, the free encyclopedia,” 2024. [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Dhaka\\_Metro\\_Rail&oldid=1214269106](https://en.wikipedia.org/w/index.php?title=Dhaka_Metro_Rail&oldid=1214269106) 32