THE UNIVERSITY OF HONG KONG

COMP4801 FINAL YEAR PROJECT

Analysing the Likelihood of Confirmation Bias in Twitter Posts

Final Report

by

Rhea Gupta

3035550731

Supervised by

Dr. Dirk Schnieders

Department of Computer Science

April 18th, 2023
Abstract

In recent years, there has been unprecedented growth in the use of social media, not only as a means to connect with family and friends and meet new people but also as a source to consume news and happenings around the world. The convenient availability of news on social media has led to a mass engagement of users with the news posts with many users commenting and reacting to these posts, expressing their views. The behaviour of users expressing biased views due to their pre-existing beliefs and ideologies is associated with a psychological phenomenon called confirmation bias. This project, thus, aims to study and analyse the existence of confirmation bias in Twitter posts. The study is conducted through a proposed solution that aims to quantify confirmation bias using statistical and machine learning models. The key deliverable of this study would be a machine learning model using both supervised and unsupervised learning techniques, that can predict whether a post display signs of confirmation bias based on the sentiments expressed by Twitter users through comments. There are three phases of this project – Phase 1 is about literature review and project proposal; Phase 2 is the backbone of the machine learning model which is the data collection and engineering stage; Phase 3 is the model building and training stage. The results from the study suggest that confirmation bias exists in the Twitter posts included in the study, as some users display signs of confirmation bias based on the comments they have written and their behaviour outside of the conversation. These users also tend to be vocal on various issues as they interact heavily with content on Twitter.
Acknowledgements

I would like to express my gratitude to everyone who has helped me bring this project to fruition. Firstly, I would like to express my sincere thanks to the Department of Computer Science for giving me this valuable opportunity.

I would like to express my gratitude to my supervisor, Dr. Dirk Schnieders for his constant guidance and support throughout this project. His immense knowledge, insightful suggestions and motivation has enabled to complete this project successfully.

Lastly, I would like to extend my thanks to my family and friends without whose encouragement and guidance this project would have been incomplete.
Table of Contents

Abstract ................................................................................................................................. i

Acknowledgements ........................................................................................................... ii

List of Figures ...................................................................................................................... v

List of Tables ......................................................................................................................... vi

List of Equations ................................................................................................................... vii

List of Abbreviations .......................................................................................................... viii

1. Introduction ...................................................................................................................... 1

1.1 Background ..................................................................................................................... 1

1.1.1 Social Media as a news provider ............................................................................. 1

1.1.2 Confirmation Bias .................................................................................................... 2

1.1.3 Echo Chambers ........................................................................................................ 2

1.2 Related Works ................................................................................................................ 3

1.2.1 Confirmation Bias .................................................................................................... 3

1.2.2 Echo Chambers ........................................................................................................ 4

1.2.3 Evaluating the studies ............................................................................................. 5

1.3 Sentiment Analysis ....................................................................................................... 5

1.3.1 Supervised Learning ............................................................................................... 6

1.3.2 Unsupervised Learning ......................................................................................... 6

1.3.3 Lexicon-based approach ....................................................................................... 7

1.3.4 Evaluating the methods ......................................................................................... 8

1.4 Motivation ...................................................................................................................... 9

1.5 Contribution ................................................................................................................... 9

1.6 Outline of the report .................................................................................................... 10

2. Methodology .................................................................................................................. 11

2.1 Application Requirements .......................................................................................... 11

2.2 Data Collection ............................................................................................................ 12

2.2.1 Collection from External Datasets ......................................................................... 12

2.2.2 Web Scraping ......................................................................................................... 12

2.2.3 Sampling bias ......................................................................................................... 13

2.3 Data Processing ........................................................................................................... 14

2.4 Sentiment Analysis Model .......................................................................................... 15

2.5 Modelling for Confirmation Bias .............................................................................. 16

2.5.1 Probabilistic Method ............................................................................................. 16

2.5.2 Proposed Criteria for Confirmation Bias .............................................................. 17

2.5.3 Verification of Results ......................................................................................... 18

2.5.4 Formal Definition of Algorithm ........................................................................... 18

2.6 Visualisation of Results .............................................................................................. 19

2.7 Summary ...................................................................................................................... 19

3. Implementation .............................................................................................................. 20

3.1 Overview ...................................................................................................................... 20
3.2 Project schedule...........................................................................................................20
3.3 Data Collection ..........................................................................................................21
  3.3.1 Collection from APIs .............................................................................................21
  3.3.2 Data Processing .....................................................................................................22
3.4 Confirmation Bias Modelling .......................................................................................23
  3.4.1 Part 1: Subjectivity-Objectivity Detection .............................................................23
  3.4.2 Part 2: Polarity Detection .......................................................................................26
  3.4.3 Part 3: Probabilistic Method to Calculate Confirmation Bias...............................28
4. Results ...........................................................................................................................29
  4.1 Case Study 1: Charlie Hebdo case ..........................................................................29
    4.1.1 Assessment ..........................................................................................................29
  4.2 Case Study 2: Twitter post by the account Kamala Harris ....................................30
    4.2.1 Assessment ..........................................................................................................30
  4.3 Case Study 3: Twitter post by the account Dr. Ian Weismann .................................31
    4.3.1 Assessment ..........................................................................................................32
5. Overall Evaluation of Results .........................................................................................34
  5.1 Data Sources ..............................................................................................................34
  5.2 Sentiment Analysis Techniques ..............................................................................35
  5.3 Identifying confirmation bias ....................................................................................36
  5.4 Verification of Results ..............................................................................................36
6. Future Works ..................................................................................................................37
  6.1 Nudge Notifications .................................................................................................37
  6.2 Visual Cues ................................................................................................................37
  6.3 Educational Pop-ups ..................................................................................................37
  6.4 Personalised Recommendations ..............................................................................37
7. Conclusion .......................................................................................................................39
References ..........................................................................................................................40
Appendix A: Case Study 1 ...............................................................................................43
Appendix B: Case Study 2 ...............................................................................................46
Appendix C: Case Study 3 ...............................................................................................48
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Main reasons for the use of social media</td>
<td>1</td>
</tr>
<tr>
<td>1.3</td>
<td>How the BERT model performs subjectivity and sentiment classification using one model</td>
<td>6</td>
</tr>
<tr>
<td>1.4</td>
<td>Dataset before and after K-Means clustering</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>Web scraping architecture</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Example of a thread of comments on Twitter</td>
<td>14</td>
</tr>
<tr>
<td>2.3</td>
<td>Example of a tree structure</td>
<td>15</td>
</tr>
<tr>
<td>3.1</td>
<td>Example of the data taken from Twitter API in JSON format</td>
<td>23</td>
</tr>
<tr>
<td>3.2</td>
<td>Comments extracted in a structured table</td>
<td>23</td>
</tr>
<tr>
<td>3.3</td>
<td>Loading of pre-trained BERT model and tokenizer</td>
<td>24</td>
</tr>
<tr>
<td>3.4</td>
<td>Results from training BERT model on the labelled dataset</td>
<td>25</td>
</tr>
<tr>
<td>3.5</td>
<td>Loading the BERT model using APIs in Python</td>
<td>25</td>
</tr>
<tr>
<td>3.6</td>
<td>Example output of VADER</td>
<td>26</td>
</tr>
<tr>
<td>4.1</td>
<td>Case study 2 tweet</td>
<td>30</td>
</tr>
<tr>
<td>4.2</td>
<td>Case study 3 tweet</td>
<td>32</td>
</tr>
</tbody>
</table>
List of Tables

Table 1.1  The labels and threshold for results from VADER algorithm as proposed by Borg and Boldt 8
Table 2.1  Application features 11
Table 2.2  Probability matrix for D and H variables 16
Table 2.3  Criteria and justification for confirmation bias 17
Table 2.4  Formal Definition of Algorithm 18
Table 3.1  Project Timeline 20
Table 3.2  Case Studies 22
Table 3.3  The important attributes and its corresponding attribute name in the table 22
Table 3.4  Score ranges for polarity detection 27
# List of Equations

<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Nickerson’s Likelihood Ratio</td>
<td>4</td>
</tr>
<tr>
<td>2.1</td>
<td>Compute conditional probabilities</td>
<td>17</td>
</tr>
</tbody>
</table>
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>BERT</td>
<td>Bidirectional Encoder Representations from Transformers</td>
</tr>
<tr>
<td>COVID-19</td>
<td>Coronavirus disease</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma separated value</td>
</tr>
<tr>
<td>EDA</td>
<td>Exploratory data analysis</td>
</tr>
<tr>
<td>HTML</td>
<td>HyperText Markup Language</td>
</tr>
<tr>
<td>HTTP</td>
<td>HyperText Transfer Protocol</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>VADER</td>
<td>Valence Aware Dictionary and Sentiment Reasoner</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Background

1.1.1 Social Media as a news provider

Today, social media usage at its all-time high. In 2021, we had over 4.4 billion people using social media platforms across the world [1]. This shows that social media attracts more than 55% of the world’s population. The unprecedented boom in social media usage can be attributed to reasons – rise in the use of smartphones and the ease of access to internet services. According to Statista, the number of smartphone users in the world in 2021 was more than 6.5 billion, accounting for 83.40% of the world’s population as compared to 49.4% in 2016 [2].

Some of the most popular and widely used social media platforms are Instagram, Facebook, Twitter, YouTube and TikTok. These platforms provide a platform for individual users and organisations such as news agencies, corporations, governmental and non-governmental organisations to create and share content for the benefit of a wider audience.

The widespread popularity and use of social media along with it acting as a source of information has enabled the general public to use it as one of the primary sources for news consumption. Moreover, even traditional media agencies such as The Economist, Bloomberg etc. have adopted to having an online presence to expand their reach beyond the conventional methods of print media.

Figure 1.1 [3] shows that reading new stories and sharing opinions with others are among the primary reasons why internet users engage on social media.

![Figure 1.1: Main reasons for using social media](image)

In addition to the traditional media houses, there has been a growing presence of alternative media entities that provide alternative news content to users. For example, in India, news sites such as The Quint, BuzzFeed and ScoopWhoop have emerged as digital media providers.
Along with reporting the latest news from around the world, they also post different opinion pieces, editorials, and commentaries on their take on the world happenings. These posts garner lot of attention from users in the form of comments and reactions.

The operations such as likes, reactions, shares and comments has enabled users to show their support for opinions and articles that resonate with them. Due to a large user base and extensive usage of social media through posts, sharing others’ contents, etc., enormous amounts of data is generated. According to Domo’s Data Never Sleeps 5.0 report, in 2017, more than 450,000 tweets were sent per minute, while more than 103 million spam emails are sent [4].

1.1.2 Confirmation Bias

With many comments, shares and reactions on a post, a vast range of opinions and viewpoints are shared. Besides the views, misinformation and fake news can also be spread through interactions on a post. The combination of diverse opinions and possible false information can lead to polarising views amongst the social media users, which in turn can cause a polarisation in the society by shaping public discourse. The behaviour of users online is often governed by their initial beliefs and ideas on a particular issue. This is associated to the psychological phenomenon called confirmation bias. Confirmation bias can be defined as the natural human tendency or inertia to seek information that confirms our initial beliefs and opinions [5].

Irrespective of their algorithms or content type, all social media websites such as Facebook, Twitter, and YouTube, serve one basic function – to bring together like-minded people based on their content preferences. Social media organisations use adaptive machine learning algorithms that track a user’s interests by assessing what one tweets or shares to keep feeding information that will keep them using the website [6]. This fosters confirmation bias as one establishes a comfort zone and keeps consuming false information.

One must ask why we need to regulate our confirmation bias while consuming information online. The answer is simple: to weaken the division and polarisation in our society caused by social media echo chambers.

1.1.3 Echo Chambers

An echo chamber is a closed and restricted environment wherein an individual’s beliefs are amplified by presenting them with information that supports their existing opinion, thereby reinforcing their confirmation bias [7]. One example of confirmation bias on Twitter can be illustrated with the use of hashtags. A Twitter user who is interested in the tweets of former US President Donald Trump can be stuck in a pro-Trump (e.g., #MAGA) or an anti-Trump (e.g., #fakepresident) echo chamber [8].
1.2 Related Works

There have been several works conducted to understand confirmation bias on certain aspects and the existence of confirmation bias in social media. Additionally, there are also works done on general sentiment analysis in social media. With the prevalence of social media, there have been much interest in this space given how social media can provide insights on various issues that are happening around the world. For instance, during the beginning of the COVID-19 pandemic, there was a study conducted to understand how the interactions of users on different social media platforms can affect the spread of misinformation regarding COVID-19 [9].

The relevant works to be discussed in this section can be broadly classified in the areas of confirmation bias and echo chambers.

1.2.1 Confirmation Bias

The past studies on confirmation bias are important in this project, as one of the objectives is to identify and quantify confirmation bias on social media comments. A number of researchers have studied confirmation bias. Some of the past works are more focused on studying confirmation bias in specific settings such as the role of confirmation bias in sharing of health content [10] on social media and confirmation bias in sequential information search [11]. Moreover, past literature also examines the broad concept of confirmation bias and statistical models to quantify it.

One such paper that was written is by Nickerson, in 1998, where he explained the phenomenon of confirmation bias [12]. This paper details the characteristics of confirmation bias as well as identifying confirmation bias using a statistical approach. It identifies that confirmation bias is the phenomenon where individuals suspecting of being bias, intentionally select evidence or give undue weight to the evidence to support their stand. The selection of evidence is also usually one-sided, either without consideration of likelihoods and evidence supporting the opposite views, or under weigh the evidence of alternative views. When opposite views are presented, bias users would also tend to behave less receptively to the alternative views [12].

Nickerson also included a statistical approach in quantifying confirmation bias, by using a Bayesian conditional probability method which defines $P(D|H)$ and $P(D|H')$, where, $P(D)$ refers to the probability of a particular observation while $P(H)$ and $P(H')$ refer to the probability of hypothesis in a particular view and an alternative view, respectively.

Hence, $P(D|H)$ refers to the probability of a particular observation if the hypothesis is true. In order to quantify confirmation bias, the paper suggested to calculate for the likelihood ratio, which is described as,
**Equation 1.1: Nickerson’s Likelihood Ratio**

\[ \text{Likelihood ratio} = \frac{P(D|H)}{P(D|H')} \]

The likelihood ratio represents the probability of a particular observation if the hypothesis is true relative to the same observation if the other hypothesis is also true. This method of quantifying for confirmation bias is used as it looks at whether there is appropriate level of attention given to the alternate hypothesis given the same observation. Hence, Nickerson provided some written characteristics of confirmation bias, and a probabilistic approach to quantify it [12].

On the other hand, there are studies conducted to analyse confirmation bias in specific settings. One of the studies was conducted by a team of scientists who aimed to investigated the role of confirmation bias in promoting social media users to share health articles online, based on the factors of health literacy and content valence of health articles [13]. The team conducted an experiment where they asked some volunteers to participate in survey questions and read selected online health articles. Through the study, the team managed to gather the results from the survey to understand the intentions of the participants in sharing the articles they have read. The team suggested from the results that there exists a relationship between confirmation bias and health literacy, on the intentions of users in sharing positive and negative health articles [13].

The above highlights some of the research works on confirmation bias, both on the broad concept of confirmation bias and confirmation bias in specific settings.

1.2.2 Echo Chambers

The next area of research is in echo chambers. A study was published recently which studied the extent of polarisation and the structure of echo chambers relating to COVID-19 discussions on Twitter in the United States [7]. The team proposed a model that can estimate user polarity from their profiles and retweets on a spectrum from left to right leaning, leveraging on a combination of content-based and network-based approaches. Content-based approach is conducted by doing sentiment analysis and topic classification of tweet data from Twitter. Network-based approach, on the other hand, is performed by analysing the networks of ‘follow’ and ‘retweet’ relationships within users on Twitter [7].

The content-based approach of sentiment analysis is performed by using a modification of the Bidirectional Encoder Representations from Transformers (BERT) model, which is a pre-trained NLP model. The team ran the pre-trained model on the user profile descriptions for sentiment analysis on user polarity. Thereafter, they also checked the topics of hashtags included in users’ profiles against a set of whitelisted left and right leaning hashtags. For the network-based approach, the team analysed the retweet behaviour of users on Twitter. This check is done by searching the media corporations that the users commonly retweet the posts...
from, against an independent media watchdog website, AllSides.com, to determine the level of media bias of these media corporations [7].

The study revealed that echo chambers exist in both the right and left leaning users, but the right-leaning users were far more connected within their echo chambers and disconnected from the users outside of their communities. This thus implies that high confirmation bias exists within the right leaning communities as the users within the communities were more receptive to information that they strongly believe in. In contrast, the left leaning and neutral users were more receptive to information from alternating views [7].

1.2.3 Evaluating the studies

The research to study the structure of echo chambers in the context of COVID-19 discussions on Twitter used a combination of content-based and network-based methods to identify the polarity of users and their activities on social media to understand the extent of echo chambers [7]. In evaluation of the approach by the team, only the content-based can be applied in this project. This is because the content-based approach consists of sentiment analysis to estimate user polarity and determine the users under which group, either left or right leaning. The method of estimating user polarity is relevant to the project as we can determine whether certain users are bias or neutral based on their expression of comments.

However, the proposed method by the team also includes a component of having to manually check the user profiles for the topics of hashtags included in their profiles against a set of whitelisted left and right leaning hashtags relating to COVID-19 discussions [7]. This component is not suitable for the project as it would require manually identifying the hashtags of each of the classes and for a wide variety of topics. Hence, besides manually checking the hashtags of user profiles, the project can thus draw some ideas from this study in the proposed method to identify for confirmation bias.

1.3 Sentiment Analysis

Sentiment Analysis is a text classification tool that analyses whether the underlying sentiment of a text is positive, negative, or neutral [14]. It uses natural language processing and machine learning algorithms to extract sentiment, opinions ad emotions from a text. There are primarily two aspects to sentiment analysis – polarity detection and subjectivity detection. Polarity detection aims to classify opinions as “positive,” “negative” or “neutral” whereas the subjectivity analysis aims to quantify the amount of personal opinion and factual information contained in a text [15] [16].

Recent works have been conducted on polarity and subjectivity detection of various kinds of text data, including product and movie reviews, as well as social media data such as comments and posts – which is more relevant to this project. These studies demonstrate various methods to perform sentiment analysis, such as using supervised and unsupervised learning algorithms, as well as lexicon-based approaches.
Lastly, there is also another task of sentiment analysis called text clustering, which is an unsupervised technique that can group sentences into various clusters based on its word embeddings [17].

Various methods can be used to perform sentiment analysis on text, such as supervised learning, unsupervised learning, and lexicon-based approach.

1.3.1 Supervised Learning

The first method is supervised learning. In this method, a labelled dataset is used to train the model to predict classes of the text in terms of positive-negative polarity and personal-factual subjectivity. A recent research published proposed a solution for using BERT on a data set for polarity and subjectivity detection, reaching up to 95% accuracy [16].

![Figure 1.3: How the BERT Model performs subjectivity and sentiment classification using one model](image)

The research performs both subjectivity classification and sentiment classification using one network of BERT model. If a comment is classified as subjective, it is run through the model again to determine the sentiment of the sentence, as shown in Figure 2.4 [16].

1.3.2 Unsupervised Learning

Unsupervised learning is a form of machine learning technique that uses algorithms to analyse and cluster unlabelled datasets by discovering hidden patterns and data grouping [18].

Paltoglou and Thelwall proposed a solution of polarity and subjectivity detection but using both an unsupervised and a lexicon-based approach. The approach made use of certain features to identify polarity and subjectivity, such as the use of emoticons, capitalisation of text and the
use of punctuation marks. The solution was assessed on datasets obtained from popular social media sites of Digg, Twitter and Myspace, where the algorithm was assessed to have performed better than some past supervised learning methods [19].

Text clustering is also an unsupervised learning technique to perform sentiment analysis. The text is grouped together into different clusters based on common words between sentences. Text clustering is performed by first creating word embeddings for all the sentences in the dataset. Word embeddings is a process where each word is mapped into an n-dimensional vector space, such that words with similar contexts will appear in closer proximity of the vector space. The word embeddings can be implemented through pre-trained NLP neural networks like Sentence Transformers [20]. Thereafter, a vector-based clustering algorithm can be applied on the word embeddings, and the algorithm returns a cluster assignment to one of the k possible clusters for each data point, where k is the number of clusters.

One of the most common techniques is the K-Means clustering technique which classifies sentences into k different clusters based on the close similarity between words of different sentences.

Figure 1.4: Dataset before and after K-Means clustering

Figure 2.5 shows how K-Means clustering works. Data points with high similarity are grouped together. The similarity is determined based on the distance between the data point and the centre of each cluster. The data point belongs to the cluster which has the shortest distance from itself.

1.3.3 Lexicon-based approach

Lexicon-based approaches are rule-based algorithms where a set of words are identified and classified into different sentiments such as positive, negative, and neutral polarity. Valence Aware Dictionary and Sentiment Reasoner (VADER) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [21]. VADER is capable of detection of polarity and intensity of emotion [21].
A recent work by Borg and Boldt proposed a solution of using the Valence Aware Dictionary for Sentiment Reasoning (VADER) library on customer’s email contents in Swedish language [22]. The pair also proposed thresholds to determine the polarity based on the score from VADER algorithm. They suggested to have five labels of polarity based on the score. The following table shows the score thresholds for each polarity label [22].

<table>
<thead>
<tr>
<th>Label</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very Negative</td>
<td>-1 to -0.65</td>
</tr>
<tr>
<td>Negative</td>
<td>-0.65 to -0.35</td>
</tr>
<tr>
<td>Neutral</td>
<td>-0.35 to 0.35</td>
</tr>
<tr>
<td>Positive</td>
<td>0.35 to 0.65</td>
</tr>
<tr>
<td>Very Positive</td>
<td>0.65 to 1</td>
</tr>
</tbody>
</table>

Another notable lexicon-based approach is TextBlob, which is also an NLP library that is available in Python [23]. TextBlob can be used in both subjectivity and polarity detection as the library can determine the results from both types of sentiment analysis [24]. The result of TextBlob is calculated using a weighted average sentiment score over all the words in the text sequences. Besides the use of words with connotations of positivity and negativity, TextBlob also uses semantic labels such as emoticons, exclamation mark and emojis to determine the subjectivity and polarity.

1.3.4 Evaluating the methods

The above methods cover some aspects of sentiment analysis that can be applied to this project, where the methods consist of supervised, unsupervised, and lexicon-based approaches in polarity and subjectivity detection, as well as unsupervised learning in text clustering.

Polarity detection is relevant to the project as the detection can estimate the user’s comments, on whether the comments are neutral or more of positive or negative sentiments, which can be used to determine the user’s hypothesis or $P(H)$ when identifying for confirmation bias. The subjectivity detection can also be used in the project to estimate the users’ comments that are subjective and objective, to identify the users that express more individual opinions than facts and vice-versa. However, subjectivity detection may not be helpful in determining...
confirmation bias but can give us an understanding of the sentiments expressed in the comments.

Text clustering method is also relevant to the project to classify the comments that belong to certain views. This is because to identify for confirmation bias, we want to obtain the comments that have similar views to calculate for $P(D)$. As such, the above suggested methods for subjectivity and polarity detection, and text clustering can be applied to the project.

1.4 Motivation

With the understanding of some characteristics of the comments made on social media posts related to news information, it is interesting to also understand the motivation of the users who make these comments. Literature suggests that humans are shaped by their prior attitudes and viewpoints when processing information [13]. Therefore, users who make biased comments on posts can be said to display signs of confirmation bias.

With the vast number of reactions and comments generated every day it might be difficult to analyse the prevalence of confirmation bias in online users – simply by browsing through the comments they post. As such, this project aims to study the overall sentiment of comments and whether polarity exists in social media posts, due to confirmation bias in online discussions.

1.5 Contribution

Digital technologies such as social media are known to be linked with social and political polarization. A key feature of social media websites is that they allow likeminded people to meet each other virtually. This allows online users to be exposed to content that amplifies their political beliefs and subsequently form social media echo chambers. Social media ranking algorithms facilitate the formation of these echo chambers by personalising and curating every user’s online experience [25].

As the geographic boundaries shrink and we get our information from biased algorithms, at a certain point, it is unavoidable to be living in a bubble. Biased users become unwilling to accept or even consider unfavourable or alternating viewpoints. Often, online users aren’t even aware of the unconscious bias that seeps into their decision-making. Therefore, it is critical for social media users to consume their information more consciously.

Through this project, we aim to contribute to making online users more informed about the existence of confirmation bias and possible filter bubbles in the Twitter posts they interact with. The ML model created will label Twitter posts as biased or unbiased based on the comments made by the users; thus, giving users an opportunity to be aware and to seek posts that highlight neutral or alternating viewpoints.
1.6 Outline of the report

The report is structured into four chapters. The first chapter highlights an overview of confirmation bias and its prevalence in social media. It also analyses past works on analysing confirmation bias in social media and sets out the goals for this project. Chapter two analyses the methodology for this project. Data collection techniques such as web scraping and data collection through APIs would be explained. Furthermore, an in-depth analysis of the machine learning techniques to be used in this project will be provided. Chapter three will detail the current progress and results or discussion obtained thus far. It will highlight the project schedule, the challenges faced while working on the project and the next steps. Chapter four is the conclusion. It sums up the project report by highlighting the key points from the other three chapters. It also highlights how this project can be improved outside the scope of the final year project.
2. Methodology

This chapter presents the technical tools and algorithms that will be used to build the model, including data collection methods and machine learning models.

2.1 Application Requirements

This proposed application is implemented to analyse sentiments from social media comments and quantify the confirmation bias from the data, if the bias exists within the comments. Hence, the application would need to have some fundamental features including from data gathering to modelling for confirmation bias. The following table describes the fundamental functional requirements for each aspect of the application.

<table>
<thead>
<tr>
<th>Features</th>
<th>Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Collection</td>
<td>1. The application should be able to extract posts and its comments from social media platforms namely, Twitter.</td>
</tr>
<tr>
<td></td>
<td>2. The application should be able to obtain the data from Twitter that contain other related information such as the timestamp and identifiers of the users and the posts.</td>
</tr>
<tr>
<td>Data Pre-processing</td>
<td>1. The application should be able to process the raw data collected into a consistent format for further steps.</td>
</tr>
<tr>
<td>Sentiment Analysis and Modelling</td>
<td>1. The application should have a model prepared to output the results of sentiment analysis and confirmation bias from the input of social media comments.</td>
</tr>
<tr>
<td>Output Visualisation</td>
<td>1. The application should be able to visualise the output in a concise format for the user to understand the results of sentiment analysis and confirmation bias.</td>
</tr>
</tbody>
</table>

The above fundamental functional requirements of the proposed application provide a direction on designing the application to fulfil those requirements. The design of each aspect is elaborated in the subsequent sections of this chapter.
2.2 Data Collection

The first aspect is data collection, where the data from social media websites needs to be gathered. As this project aims to analyse social media posts data on news articles and commentaries, the extracted data should be taken from the relevant sources, which are accounts belonging to the news providers. Most of the well-known news providers in Hong Kong such as South China Morning Post and Hong Kong Free Press, as well as global news providers such as Bloomberg, Reuters and The New York Times have presences in the notable social media platforms like Facebook, Twitter, and Instagram. For the purpose of this project, we will be focusing on extracting data from Twitter.

There are two ways of performing data collection, either by using external datasets, or by web scraping the data.

2.2.1 Collection from External Datasets

The first method of data collection is by looking for external social media datasets collected by past researchers. These datasets are usually collected to serve certain research objectives, and some of these datasets are labelled and can be used in machine learning applications. They can be downloaded and stored in files, can be read using Python libraries. These datasets, hence, do not require the calling of API endpoints.

**Advantages:** We are able to collect the data without having to make API calls to the social media platforms; which usually would require some form of access control through an account login and some authentication bearer tokens.

**Disadvantages:** The collected datasets from past research are likely to be outdated, hence, may not be relevant to present day. Moreover, these datasets are collected keeping in mind certain objectives of the studies which might not suit the needs of this study.

2.2.2 Web Scraping

The second method is by web scraping the data from the social media websites using APIs. Web scraping is an automatic method to obtain substantial amounts of data from a website by extracting the underlying HTML code and, with it, the data stored in the database [26]. Most of the extracted data is unstructured and needs to be converted to a structured format. Python is the most popular language for web scraping as it has many libraries specifically designed for web scraping [27].

A web scraper is a program that extracts large amounts of data from a target website by sending HTTP requests to the website server [28]. It parses content from both publicly accessible data and internal APIs of the website.
The following steps demonstrate a general web scraping process, as shown in Figure 2.1:

i. Identify target websites for the scraping script to attack.
ii. Make the request to the website URLs to extract the HTML of the page
iii. Use HTML tag locators to find the required content on the page
iv. Save the data in a structured data format e.g., CSV, JSON or spreadsheet.

**Advantages:** It provides the flexibility to obtain most recent data directly from the platforms so that the extracted data is more up to date with the recent developments that we want to analyse. A web scraper is more flexible and highly scalable approach as it can developed to specifically cater to the needs of this project. Specific metadata information such as unique identifier numbers can be easily extracted.

**Disadvantages:** As websites constantly change their structure, it is important to ensure that the scraper does not break. It requires regular maintenance to ensure that the data collection pipeline stays clean [29].

For this project, web-scraping will be used for data collection and will form the backbone of the ML model. The primary reason why this method is chosen is because it provides more flexibility while extracting the data. The dataset needs to be of specific format and needs to contain certain features such as user information, comments in the form of text and post details. As past datasets might have restricted access to such data, it is more feasible to build a custom web scraper that facilitates this project.

2.2.3 Sampling bias

Sampling bias is a type of machine learning bias which can skew the results produced by a model. This bias occurs during the initial stages of development, namely data collection. Data is the core to any machine learning algorithm. If the data provided to the model is incorrect or
skewed or has faulty data points, it can lead to inaccurate results for the given sample set [30].

For the purpose of this project, it is acknowledged that sampling bias might cause skewed or inaccurate results by the machine learning bias. Caution will be taken when sampling the data to minimise the effects of sampling bias.

A few proposed precautions are:
1. Choosing a large dataset
2. Using a diverse range of Twitter posts
3. Using random sampling to form the dataset
4. Testing and finetuning the parameters based on the model performance.

2.3 Data Processing

Next, the collected raw data has to be pre-processed into a consistent format. The raw dataset collected might have information that is not relevant. For the purpose of this project, only a few important attributes are relevant. These attributes can be obtained and stored in a structured tabular format.

The following are the important attributes:

1. Unique identifier for users who commented on the post.
2. Account details of the source of main post.
3. Thread information based on replies on comments.
4. Timestamp for each comment.
5. The comment in text sequences.

The user’s unique identifier is relevant to this project because in the social media context, the comments can be replied by other users that form a discussion of a particular issue. As the project aims to find out on the existence of confirmation bias within these discussions, it is relevant to consider the relationship of replied comments with the main comment. Figure 2.2 shows an example of chain of comments that form as users reply to each other.

![Figure 2.2: Example of a thread of comments obtained from Twitter][31]
With the relationship between the primary comment (or the root node) and its child comments (or the child nodes), the comments can then be organised in a tree structure as illustrated below in Figure 2.3. A reply to a parent or a root comment becomes the child node which can further have more child nodes.

![Figure 2.3: Example of a comments tree structure extracted from Case Study 1](image)

### 2.4 Sentiment Analysis Model

After preparing the data for analysis, the next process is to perform the sentiment analysis. As identified in the proposed solution by Jiang, Ren and Ferrara, the characteristics of echo chambers and polarity of users can be identified using the content-based approach of sentiment analysis [7]. Hence, this application can conduct sentiment analysis using a combination of methods proposed in section 1.3, to identify for comments that exhibit confirmation bias. The sentiment analysis, thus, includes polarity and subjectivity detection of the comments, which can be achieved through separate models for both polarity and subjectivity detection. Additionally, the application can also include text clustering to classify the comments into two different classes, where one is that of supportive comments while the other class is that of unsupportive comments.

Section 1.3 also includes various approaches of sentiment analysis, namely, using the supervised learning, unsupervised learning, or lexicon-based approaches. The papers referenced include details on the specific methods used, such as the BERT model for supervised learning, the VADER and TextBlob libraries for lexicon-based approach and K-Means clustering for unsupervised learning, which can be adopted in this project too. Like TextBlob, the output from subjectivity detection is a probability value in a range of 0 to 1, where a value close to 0 refers to an objective sentiment while, a value close to 1 refers to a subjective sentiment. Like VADER, the output from polarity detection is a decimal value in a range of -1 to 1, where a value close to -1 refers to sentiment of negative polarity while, a value close to 1 refers to sentiment of positive polarity.
Hence, for this project, these specific methods would be used on the social media comments and assess which of the specific methods are more suitable. The assessment of the methods is performed in later chapters.

2.5 Modelling for Confirmation Bias

The next aspect is of designing the model for identifying and quantifying confirmation bias. To identify for confirmation bias, we want to know the users that are potentially bias as well as an assessment of biasness in the entire conversation.

2.5.1 Probabilistic Method

The confirmation bias can be quantified using the proposed probabilistic method suggested by Nickerson. The application can include a function to calculate the level of confirmation bias within the conversation, by using the probabilistic method of likelihood ratio as mentioned in Section 1.2 [12].

The results can be computed by considering the probability of hypotheses, \( P(H) \) and \( P(H') \), through computing the probability of the number of comments that are of positive and negative sentiment respectively. The number of sentiments that are of positive and negative polarity can be identified during the sentiment analysis process through polarity detection. As for the probability of observation, \( P(D) \), the application would consider the variable \( D \) as the number of comments that are in favour of the view set out in the particular news article or commentary, while \( D' \) is the number of comments that are not in favour of the view. The variables \( D \) and \( D' \) can be computed by taking the clusters of comments that represent the supportive and unsupportive views. After classifying the number of comments based on the two clusters, we can then derive \( P(D) \).

Hence, the application can generate a probability matrix like the table below to calculate the conditional probability. From the matrix table, a, b, c and d are some probability values less than 1, where \( a + b + c + d = 1 \).

<table>
<thead>
<tr>
<th></th>
<th>( P(H) )</th>
<th>( P(H') )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(D) )</td>
<td>a</td>
<td>b</td>
<td>( a + b )</td>
</tr>
<tr>
<td>( P(D') )</td>
<td>c</td>
<td>d</td>
<td>( c + d )</td>
</tr>
<tr>
<td>Total</td>
<td>( a + c )</td>
<td>( b + d )</td>
<td>1</td>
</tr>
</tbody>
</table>

Based on these values, the application can compute for \( P(D|H) \) and \( P(D|H') \). The following formulas show the computation of both probabilities.
Equation 2.1: Compute conditional probabilities

\[
P(D|H) = \frac{P(D \cap H)}{P(H)} = \frac{a}{a + c} \quad \text{and} \quad P(D|H') = \frac{P(D \cap H')}{P(H')} = \frac{b}{b + d}
\]

The statistical method outputs a probability value in a range of 0 to 1, based on the computation of \(\frac{P(D|H)}{P(D|H')}\). A value of close to 0 refers to very high existence of confirmation bias while a value close to 1 refers to a low existence of confirmation bias.

2.5.2 Proposed Criteria for Confirmation Bias

With the above proposed methods to identify for confirmation bias, the following table is the proposed criteria to determine for confirmation bias, where we want to identify the users who are potentially bias and the overall biasness of the conversation.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polarity of the comments to determine potentially bias users</td>
<td>If the comment is of high in polarity, regardless of whether it is negative or positive polarity, there is a tendency for echo chambers and confirmation bias. This is because, if the comment is high in polarity, it may imply that the users may be making comments that are not neutral and are one-sided to reinforce their own beliefs in the discussion, without considering other beliefs [12]. In contrast, a neutral view would be where the comment provides a balanced viewpoint.</td>
</tr>
<tr>
<td>Likelihood ratio of (\frac{P(D</td>
<td>H)}{P(D</td>
</tr>
</tbody>
</table>
2.5.3 Verification of Results

The process of performing sentiment analysis would help us identify the users that are potentially biased, while computing the statistical method of likelihood ratio can allow us to understand the biasness of the entire conversation. As such, after the application estimates the biased users, the final step is to verify that the users are indeed bias.

The verification can be performed in a two-step approach. The first step is to manually check the comments to assess if it is bias, and the second step is to apply the network-based approach as proposed in Section 1.2 [7], where we can analyse the retweet behaviour of the users who are potentially biased. Specifically in the second step, we can look at the users’ retweets, likes and replies on other tweets to determine if a particular user is consistent with his/her views on the topic, and whether the user also considers the alternative view. This second step is also verified manually.

2.5.4 Formal Definition of Algorithm

In summary, the following steps illustrate formally how the algorithm can identify and quantify for confirmation bias in the social media comments.

*Table 2.4: Formal Definition of Algorithm*

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Select an original post from a news provider or influential account on Twitter and obtain the unique identifier number of the post. From the post, process the original raw data and collect key attributes and details. Thereafter, save the data as a structured tabular format and save locally.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 2</td>
<td>From the obtained social media comments, clean the original text sequences by removing punctuation marks and related symbols unique to the social media platform (such as the username starting with ‘@’ in Twitter).</td>
</tr>
<tr>
<td>Step 3</td>
<td>Perform polarity and subjectivity detection on the cleaned text sequences and obtain the corresponding scores for each comment. Based on the scores, estimate and determine the polarity and subjectivity classes of each comment.</td>
</tr>
<tr>
<td>Step 4</td>
<td>Perform text clustering on the comments to estimate and determine the two classes of supportive and unsupportive comments.</td>
</tr>
<tr>
<td>Step 5</td>
<td>Based on the determination of polarity detection and text clustering, compute the likelihood ratio [12].</td>
</tr>
</tbody>
</table>
2.6 Visualisation of Results

The results from the sentiment analysis and modelling of confirmation bias are then collated and presented in a tabular format, so that we can look at each comment and its corresponding sentiment analysis scores.

2.7 Summary

This chapter outlined the major technologies and platforms to be used for this project. Web-scraping and structured data formats were explained with illustrations. The algorithms to be used for the machine learning model were explained and evaluated. The next chapter will highlight the implementation of the project.
3. Implementation

3.1 Overview

This chapter highlights the implementation process of the project. The sections below detail the project timeline challenges faced and the limitations of the project.

3.2 Project schedule

Table 3.2 shows the project schedule. Phase 1 deliverables which consisted of a detailed project plan and project website have been completed and submitted. Currently, the project is on track with the schedule mentioned in Table 3.2 and is in the data collection stage.

The first phase of the project was the ideation. This stage included the Phase 1 deliverables and literature review. After completion of this phase, we move on to Phase 2, which is the elaboration stage. It includes getting the data for the model. The building of the web-scraper is the backbone of this project, as stated in 2.1.2, hence, requires thorough research and implementation. Data is the key element of any machine learning model and determines its success. The building of the web-scraper is the backbone of this project, as stated in 2.1.2, hence, requires thorough research and implementation. The final phase – Phase 3 – is the conclusion stage. It consists of the model training using the processed data from Phase 2, code review and final deliverables submission.

Table 3.1: Project Timeline

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 2022 – October 2022</td>
<td><strong>Phase 1 – Ideation and research</strong></td>
</tr>
<tr>
<td></td>
<td>- Background research on confirmation bias and existing studies</td>
</tr>
<tr>
<td></td>
<td>- Familiarization with sentiment analysis and web-scraping techniques</td>
</tr>
<tr>
<td>2nd October 2022</td>
<td><strong>Key Milestone: Phase 1 Deliverables</strong></td>
</tr>
<tr>
<td></td>
<td>- Detailed Project Plan</td>
</tr>
<tr>
<td></td>
<td>- Project web page</td>
</tr>
<tr>
<td>November 2022 – January 2023</td>
<td><strong>Phase 2 – Implementation</strong></td>
</tr>
<tr>
<td></td>
<td>- Web-scraper building</td>
</tr>
<tr>
<td></td>
<td>- Data cleaning and processing</td>
</tr>
<tr>
<td>18th January 2023</td>
<td><strong>Key Milestone: Interim Deliverables</strong></td>
</tr>
<tr>
<td></td>
<td>- Interim Report</td>
</tr>
<tr>
<td></td>
<td>- Interim Presentation</td>
</tr>
<tr>
<td>February 2023 – March 2023</td>
<td><strong>Phase 2 – Implementation (cont’d)</strong></td>
</tr>
</tbody>
</table>
3.3 Data Collection

3.3.1 Collection from APIs

The first implementation is the data collection component. As mentioned in Section 2.1, a web scraping model using Twitter API and Python BeautifulSoup library is used to collect the dataset. Twitter is chosen as the social media platform of preference because it has multiple available APIs that can be utilised to access content and meta-content, such as information on an individual user and information pertaining to a specific tweet. Additionally, the Twitter APIs are easy to use and the documentation for using the APIs is available as well. Twitter is also widely used by many users globally, with an estimated 360 million users [3]. Twitter API v2 were used to form endpoints to collect the data in the required format.

Although the social media platforms such as Facebook and Instagram also have large amounts of data available for mining, the APIs of both the platforms are unfortunately not as easily accessible as compared to Twitter due to its relatively stricter access. Hence, for this project, the choice of social media platform is Twitter.

Another source of data for this project is from social media data collected previously to support past research works. A relevant dataset for this project is the PHEME dataset for rumour detection and veracity classification, a research work conducted by a team comprising of Kochkina, Liakata and Zubiaga [32]. In the study, the team collected a labelled dataset of tweets from Twitter – which are extracted tweets posted during breaking news – that consists of two classes: rumours and non-rumours. The tweets are organised according to the events of breaking news, covering events such as the Charlie Hebdo shooting in January 2015 and the crash of Germanwings Flight 9525 in March 2015 [32].

This dataset is relevant to the project because the data is collected based on news events and it is taken from Twitter, hence, the data is similar to the types of tweets that this project aims to analyse. Therefore, this dataset is relevant to the project and is used in the project as part of data collection.
The following table shows the selected posts and their respective sources chosen for model analysis.

**Table 3.2: Case studies**

<table>
<thead>
<tr>
<th>Case Study</th>
<th>Post</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Charlie Hebdo shooting</td>
<td>PHEME Dataset</td>
</tr>
<tr>
<td>2</td>
<td>“It’s time for Congress to codify the protections of Roe into federal law.” – Kamala Harris</td>
<td>Twitter</td>
</tr>
<tr>
<td>3</td>
<td>“She asked her OB-GYN what her options were when she learned of the fatal diagnosis. &quot;Well, because of the new #abortion law, you don't have any options. You have to go on with your pregnancy.&quot; Their baby with anencephaly died within a few hours of birth.” – Dr. Ian Weismann</td>
<td>Twitter</td>
</tr>
</tbody>
</table>

3.3.2 Data Processing

After the data collection is completed, the next part is processing of the original raw data. The raw data obtained from the Twitter APIs is in the JSON format. The JSON format is a type of key-value data model where there is a key which defines the values. This format is similar to the dictionary data type in Python programming language, which thus allows an easier manipulation of the JSON object with Python.

Some important attributes as identified earlier in Section 2.1, are: the unique identifiers of the specific social media post and users who are part of the discussions, the timestamp of each comment, and the actual comments in text sequences. Based on the key-value pairs of data from the JSON object, the application can extract the values of these important attributes using Python and store the values in a structured tabular format. The structured table will thus have the attributes as seen from the following table below.

**Table 3.3: The important attributes and its corresponding attribute name in the table**

<table>
<thead>
<tr>
<th>Parameters in the table</th>
<th>Important attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Unique identifier of user who made the comment</td>
</tr>
<tr>
<td>reply_to</td>
<td>Unique identifier of the users replying to the comment</td>
</tr>
<tr>
<td>timestamp</td>
<td>Timestamp of each comment</td>
</tr>
<tr>
<td>comment</td>
<td>The comment in text sequences</td>
</tr>
</tbody>
</table>
Each record of the table corresponds to each comment of the social media post, such that the collected data has been transformed from its raw format in JSON to a structured table. The following figure is an example of the structured table.

<table>
<thead>
<tr>
<th>id</th>
<th>timestamp</th>
<th>reply_to</th>
<th>comment</th>
<th>social_media</th>
<th>conversation_id</th>
<th>head_id</th>
<th>user_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2023-04-17T20:07:28.000Z</td>
<td>@KamalaHarris @TheDemocrats <a href="https://t.co/yhM7x">https://t.co/yhM7x</a>...</td>
<td>Twitter 164363812780479631 30354991 15951836665602908160</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2023-04-17T15:26:25.000Z</td>
<td>@KamalaHarris <a href="https://t.co/0slgYdYpa">https://t.co/0slgYdYpa</a></td>
<td>Twitter 164363812780479631 30354991 131304350542193665</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2023-04-17T10:09:01.000Z</td>
<td>@KamalaHarris <a href="https://t.co/D2Vzh">https://t.co/D2Vzh</a>...</td>
<td>Twitter 164363812780479631 30354991 1599158396856908160</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2023-04-17T01:52:37.000Z</td>
<td>@KamalaHarris Pull your head out of your ass</td>
<td>Twitter 164363812780479631 30354991 1643706577436232924</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2023-04-17T23:22:06.000Z</td>
<td>@KamalaHarris It's time for you to leave office</td>
<td>Twitter 164363812780479631 30354991 1313480587630557906</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2023-04-17T23:12:31.000Z</td>
<td>@KamalaHarris It's been time. But later it bet...</td>
<td>Twitter 164363812780479631 30354991 16285344882444910593</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2023-04-17T11:01:38.08.000Z</td>
<td>@KamalaHarris No that's just stupidity</td>
<td>Twitter 164363812780479631 30354991 19740405944141401090</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2023-04-09T12:46:29.000Z</td>
<td>@KamalaHarris It time for you to start doing u...</td>
<td>Twitter 164363812780479631 30354991 1523122092985723284</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2023-04-09T02:14:14.000Z</td>
<td>@KamalaHarris Are you this dumb? It is a stat...</td>
<td>Twitter 164363812780479631 30354991 1820523528</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2023-04-09T03:42:02.000Z</td>
<td>@KamalaHarris Go to hell you devil woman!!!</td>
<td>Twitter 164363812780479631 30354991 15740405944141401090</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Machine Learning Method with BERT

The BERT model is implemented using Transformers, a deep learning framework developed by Hugging Face [33]. The Transformers framework is available as a package in Python, which allows the execution of any available deep learning model [33]. There are many pre-trained models available from Hugging Face that can be executed with Transformers, BERT is one of them.

Implementing a model from Transformers in Python is also very convenient, as the model can be loaded into the machine by calling the API from the Transformers library and specify the name of the pre-trained model for use in the training. The following figure below illustrates an example of loading a pre-trained model and the tokenizer from Transformers [34].

![Figure 3.3: Loading of pre-trained BERT model and tokenizer](image)

The tokenizer is saved as the pre-trained model tokenizer is required for the model training. The tokenizer encodes the text sequences into numerical values, in order for the model to identify the sequences through the token values.

For this project, the specific BERT model to use is the ‘bert-base-uncased’ model [34]. This model is trained with supervised learning on raw texts from a large corpus of English text data – where the data is taken from English Wikipedia and BooksCorpus, a dataset consisting of 11,038 unpublished books. There are two objectives in training this BERT model, namely, masked learning modelling and next sentence prediction [34].

Training the Model with BERT

The BERT model can be used in subjectivity detection by training the model against a labelled dataset as implemented by the team comprising of Satapathy, Pardeshi and Cambria [16]. In this project, the model is trained against the same dataset used by the team, which is a labelled dataset of subjective and objective movie reviews from Cornell University [35] [36]. The dataset comes in two separate directories – one folder for the objective reviews and the other folder for the subjective reviews.

Hence, further pre-processing of the data is performed such that the data is put together and the subjective reviews are represented by the value of 1 while the objective reviews are represented by the value of 0. The dataset is then split to an 80% train set and 20% validation set ratio. After training the model for three epochs, the model achieved 96% accuracy on the validation set.
set. Finally, the trained model is saved locally for deployment. The following figure below shows the accuracy and loss results after training for each epoch.

![Figure 3.4: Results from training BERT model on the labelled dataset](image)

**Model Deployment**

After training the BERT model, the model can be deployed on the text sequences of social media comments to determine whether the text is subjective or objective. Previously, the trained model and tokenizer are saved locally. Hence, the model and tokenizer can be conveniently deployed by using the APIs from Transformers to load the saved model and tokenizer from the local directory [34]. The following figure below illustrates an example of loading the model and tokenizer from the directory. After loading the model, it can be used to assess the text sequences for subjectivity-objectivity detection.

![Figure 3.5: Loading the BERT model using APIs in Python](image)

**Lexicon-based approach with TextBlob**

The next approach for subjectivity-objectivity detection is the lexicon-based approach with the TextBlob library. TextBlob can be implemented by calling the Python library and pass the text sequences as an input to the TextBlob function [23]. Thereafter, the subjectivity score can be returned as the output, which the score is a value in the range of 0 to 1.
Interpreting the results

After obtaining the score from subjectivity detection, the application determines the text sequences that constitute for an objective or subjective sentiment. The score of above 0.5 would mean that the sequence is subjective, while score of below 0.5 implies that the sequence is of objective sentiment.

Evaluation

As both methods compute and estimate subjectivity differently, we would get different scores for both the methods. The worst-case scenario is where the scores of both methods classify a comment in two different classes (i.e., method A classifies it as subjective and method B classifies it as objective). Hence, to further determine overall subjectivity, the application classifies a comment as subjective if both the methods return scores of above 0.5, while a comment is classified as objective if both the methods return scores of below 0.5. If both the scores classify a comment in different classes, the comment is indicated as “Unknown”.

3.4.2 Part 2: Polarity Detection

The next part of sentiment analysis is polarity detection. In this project, we use only the lexicon-based approach, by using the VADER and TextBlob libraries.

Lexicon-based approach using TextBlob

One method for polarity detection is with the TextBlob library. Similar to subjectivity detection, TextBlob is implemented in Python by calling the library and passing the text sequences as an input to the function \[23\]. We receive the polarity score as the output.

Lexicon-based approach with VADER

Another method for polarity detection is with the VADER library. Like TextBlob, VADER can be implemented in Python by calling the library, vaderSentiment, and using a new class instance SentimentIntensityAnalyzer \[37\]. From the class instance, it can be used to call the function, polarity_scores, and pass the text sequence as the input to the function \[37\]. The result from this function is a Python dictionary of four keys, namely, pos, neg, neu and compound. The following figure below shows an example output of the function.

```python
string = 'Give us believable reasons why we need mayors And what\'s with that ridiculous high salaries for mp'
getSentimentalResults(sid_obj, string)
(0.10888888888888888, 0.68, {'neg': 0.135, 'neu': 0.065, 'pos': 0.0}, -0.3012)
```

**Figure 3.6: Example output of VADER**

There are three keys in the output – pos, neg and neu – which constitute the probability values of the text sequence to be of positive, negative, and neutral polarities respectively. The values from each of the three keys will add up to 1. The key with the maximum probability value is
the most likely polarity as predicted by VADER. In the case of the above example, the neu key has the highest probability value of 0.865, hence the sentiment is likely to be of neutral polarity. Lastly, there is also a compound key, which is a compound score considering results from the other three keys. The compound score is in the range of -1 to 1; a value closer to -1 means sentiment of negative polarity, while value close to 1 means sentiment of positive polarity. From the above example, the compound score is -0.3612 which shows that the sentiment is likely to be of negative polarity.

Interpreting the results

After obtaining the score from polarity detection, the application determines the text sequences that constitute a positive, neutral or negative sentiment. The score interpretation is adopted from the proposed method by Borg and Boldt as outlined in Section 1.3 [22]. Hence, the following table illustrates the polarity classification and its corresponding score range to be used in the project.

<table>
<thead>
<tr>
<th>Label</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>-1 to -0.35</td>
</tr>
<tr>
<td>Neutral</td>
<td>-0.35 to 0.35</td>
</tr>
<tr>
<td>Positive</td>
<td>0.35 to 1</td>
</tr>
</tbody>
</table>

Evaluation

Similar to subjectivity detection, there can be different scores from both methods of polarity detection as both methods compute differently. The worst-case scenario is also where the scores of both methods classifies a comment in different classes (i.e., method A classifies it as positive polarity and method B classifies it as negative polarity). As such, to further determine overall polarity, the following are the 4 cases of polarity and its corresponding criteria.

- A comment is positive in polarity if the scores of both methods are either above 0.35, or one score is above 0.35 while the other score is above 0.
- A comment is negative in polarity if the scores of both methods are either below -0.35, or one score is below -0.35 while the other score is below 0.
- A comment is neutral in polarity if both the scores are in the range of -0.35 and 0.35.
- Otherwise, the comment is indicated as “Unknown” (in the case where both the scores classify a comment in different classes).
3.4.3 Part 3: Probabilistic Method to Calculate Confirmation Bias

To determine the results of confirmation bias through the probability method, we perform various calculations. Firstly, the matrix table is generated for the probability values. The results from polarity detection performed earlier are used to compute the probability values of \( P(H) \) and \( P(H') \). \( P(H) \) is computed by taking the total number of comments that are of positive polarity sentiment over the number of all comments, while \( P(H') \) is computed by taking the total number of comments that are of negative polarity sentiment over the number of all comments.

Results from Text Clustering

In order to compute the probability values of \( P(D) \) and \( P(D') \), an additional process of text clustering is performed. Text clustering is useful in this project as it enables the computation of clusters of comments that of the similar observations. The text clustering process is performed from implementing word embeddings first, where the word embeddings are implemented by using the Sentence Transformers library in Python, which makes use of one of the pre-trained networks to create the embeddings [20].

In this case, the pre-trained network, 'all-MiniLM-L6-v2', is used to create the embeddings. After the word embeddings have been implemented, the clustering algorithm is performed on the word embeddings [38]. In this project, the clustering algorithm used is the K-Means clustering algorithm to assign and classify the sentences within the clusters. The K-Means algorithm is also implemented in Python through the Scikit-learn library.

Interpreting the results

The likelihood ratio computed is based on the entire conversation of comments of a specific tweet, where the computed likelihood ratio is a decimal value in the range of 0 to 1. A value closer to 0 refers to higher confirmation bias as there is higher probability of a particular hypothesis given a particular observation in relative to the alternative hypothesis given the same observation. In contrast, a value closer to 1 refers to lower confirmation bias as the probability of a hypothesis given an observation is relatively close to the probability of the alternative hypothesis given the same observation.
4. Results

4.1 Case Study 1: Charlie Hebdo case

Context

On January 2015, there were reported terror attacks at the office of Charlie Hebdo newspaper in Paris, France. Additionally, the suspects of the terror attacks also struck a kosher grocery store and the Paris suburb of Montrouge. The terror attack killed 17 individuals [39].

About the Post

The tweet was posted on 9th January 2015 by SkyNews, and it is about a report on police arresting some young people on scooters near a kosher store that is close to the sites of attacks at Paris. The post included a link to the article.

Source: PHEME dataset based on Twitter posts and comments

Results

Confirmation bias score of likelihood ratio: 0.179
Number of potentially biased comments: 12
Total number of comments: 23

4.1.1 Assessment

Looking through the comments in this post, the text clustering algorithm does not seem to have separated the comments into differing views on the topic. This is because, the text clusters were split into two classes, namely, comments that contain terms relating to police and arrests, while the other cluster have comments that do not contain these terms.

As for polarity detection, it works relatively well in identifying for different sentiments. For instance, a comment reads “WTF”, which is an offensive term, and another comment that contains the phrase “filthy children killers army” were labelled as negative polarity, while comments like “Israel thanks you for another $100. Haha the irony of you helping Israelis is beautiful” and “I'm donating $100 to Israeli charities for every tweet you make today on top of the usual amount” were labelled as positive polarity, even though the comments may not seem to be related to the discussion of the main post, but because these comments contain positive terms like “beautiful” and “donate”, the comments seem to be in positive polarity. This shows that the polarity detection techniques did not consider the context of the comment and is just estimating the polarity based on some key words.

In terms of subjectivity and objectivity, the algorithm also manages to classify some comments correctly. For instance, a comment reads “what did these youngsters do to be arrested?” was classified as an objective comment perhaps because the comment is phrased into a question.
Another comment reads “Tensions are high, blah blah” was labelled as subjective because the comment seems to be more imbalanced.

However, not all the comments are classified correctly as some of them do not make sense or are irrelevant to the post. For example, there are a few comments that are retweets, which are retweeting the exact same tweet from the main post. As such, the results from sentiment analysis hence does not make sense for these kinds of comments.

From the results, it is observed that the entire discussion thread has a confirmation bias likelihood at 0.179. The low value could be attributed to that fact that about 17% of the comments are classified as positive polarity in contrast to about 9% of the comments classified as negative polarity. This difference in number of comments in both polarity classes could be the cause of the low value in likelihood ratio. This is because the probability value of P(H) is more than P(H’).

In terms of the performance in sentiment analysis, it seems that the polarity and subjectivity detection had performed well in identifying comments that are of positive and negative sentiment based on some of the words. However, the text clustering did not classify the comments based on the different views of the topic as we intended in the project.

The full results of this analysis can be found in Appendix A.

4.2 Case Study 2: Twitter post by the account Kamala Harris

Tweet

![Kamala Harris tweet](image)

*Figure 4.1: Case study 2 tweet*

**Source:** Twitter

**Results**

**Confirmation bias score of likelihood ratio:** 0.53

**Number of potentially biased comments:** 4

**Total number of comments:** 12

4.2.1 Assessment

Looking through the comments, the text clustering does not seem to be able to capture the differing views of the topic. This is because, one of the 2 clusters is assigned to comments that
include the links in them. In contrast, the other cluster has comments that do not contain any external links. On cleaning the data further to drop comments that contain external links, we notice better results in clustering data. For example, comments containing words such “yes” or “thanks” are clustered together while, all others are clustered together. This shows that the clustering techniques did not consider the context of the comment and is just estimating the clusters based on some key words.

As for the polarity detection, the models performed relatively well in identifying for polarity. This is because the model could detect certain comments that are of positive, negative, and neutral polarity. For example, a comment included the terms, “dumb” and “issue” was classified as negative sentiment while the comment with terms like “better” is classified as positive polarity. However, the polarity detection in these comments does not consider the context in determining the polarity. For example, one comment reads “The Supreme Court says otherwise.”, and is classified as a positive sentiment, but this comment seems to be anti-abortion law and hence should be a negative sentiment. Additionally, comparing the results from TextBlob and VADER, both the techniques give similar results.

Based on the above results, we can conclude that there exist some users who are biased. This is because there are some users who are vocal on the similar topic as the main conversation in their retweets, likes and replies, and also continue to support their views rather than accepting alternative views. Additionally, there are also users who included evidence in their comments in the form of external links to further strengthen their views. In terms of the overall bias, the likelihood ratio is high at 0.53, which suggests that there is relatively lower level of bias in the entire conversation. The high likelihood ratio could be attributed to the fact that while there are more comments classified as negative polarity than the comments classified as positive polarity, there is still a balance between the hypotheses.

With regards to the performance, the performance of text clustering did not perform as expected to distinguish the users in differing views, because of the lack of contextual understanding by the model. However, the polarity detection models performed relatively well, and both VADER and TextBlob performed well.

The full results of this analysis can be found in Appendix B.

4.3 Case Study 3: Twitter post by the account Dr. Ian Weismann

Tweet
Results

**Confirmation bias score of likelihood ratio:** 0.261
**Number of potentially biased comments:** 2
**Total number of comments:** 10

4.3.1 Assessment

Evaluating the performance of sentiment analysis in the comments, the polarity detection methods have classified some comments correctly in positive and negative sentiment. Comments such as ‘This is heart-breaking. No one should have to go through that. These abortion laws are so cruel.’ are classified as strongly negative with TextBlob value as -1. Due to the nature of the post, no comments have been classified as positive and on manual evaluation of the comments, it turns out to be true. However, comments that should have been classified as negative have been classified as neutral by both VADER and TextBlob. This could be attributed to use of subtle words in the comments and passive language.

As for the output of text clustering, like the previous case studies, the output did not classify the comments based on the differing views, but rather based on the common terms. One of the clusters have comments referring to terms such as “horrendous”, “tyranny” and a broken heart emoji while the other cluster contains all the remaining comments. These comments have key words such as “heartbroken” and “unconstitutional” and should be clustered with the other cluster. This unexpected classification of clustering could be attributed to the lack of context.
provided to the algorithm in classifying the comments to the supportive and unsupportive views.

Based on the above results, we can conclude that there exist some users who are biased, as we can see that they are very vocal about their views on this topic. Moreover, the language used by users in this post suggests that some users are strongly opinionated about the topic of discussion and do not hesitate to use strong language to show it. Even though the polarity detection did not correctly predict all the comments, and the text clustering did not classify the comments in the different views, we can still do the analysis based on the already identified potentially biased users. The verification of these users also allows us to further conclude the existence of confirmation bias within these users.

The likelihood ratio is at 0.261, which is quite low and implies higher confirmation bias. This can be attributed to the fact that most comments are classified as neutral or negative polarity, thus the conditional probability values of \( P(D|H') \) may be much higher, making this post highly biased towards the negative opinion.

It is important to note that the results for this case study cannot be used to make accurate judgements about the post as the model does not perform as expected. This could be due to various reasons mentioned above.

The full results of this analysis can be found in Appendix C.
5. Overall Evaluation of Results

After evaluating the performance of the application in Chapter 4, this chapter focuses on the overall evaluation of the project by outlining some strengths, limitations and challenges in different aspects of the project.

5.1 Data Sources

In this project, we obtain social media data from two sources, namely from Twitter APIs and the PHEME dataset. Both sources provide large amounts of data. However, as we see from Case Study 1, the data from PHEME dataset can be outdated as it was collected previously. Hence, some tweets that exist in the PHEME dataset may no longer exist on Twitter now. Moreover, some comments could have also been posted after the collection process, hence the dataset does not have these tweets. Therefore, if we would like to do an in-depth analysis on the users, the results would be lacking.

On the other hand, collecting data from Twitter APIs gives us more up-to-date tweets and we can perform in-depth analysis of biasness of the users. Additionally, the Twitter API documentation is thorough and online support is available for the same. However, Twitter APIs are restrictive in nature as they impose limits on the amount of data once can access given the type of Developer account permissions. For this project, we are using the Twitter Essential Access Account. However, this level only allows to extract tweets up to 7 days prior from the request. To search the Twitter archive, Elevated access is required. However, it is a hard to obtain permission level and is time consuming. One drawback of this limit is the inability to collect the comments of older tweets, which would be advantageous for the project as we can consider historic tweets.

As mentioned in Section 2.2, sampling bias is one of the major limitations of the data collection. As the Twitter dataset is large, it not feasible to analyse all or even a majority of posts for this study. Therefore, for the scope of this project, 2 Twitter posts and one external dataset have been selected at random which are from different sources such as news providers and personal accounts. These posts have been chosen based on the number of likes, shares, comments they have received. Therefore, it is acknowledged that sampling bias might be present in the dataset.

Another limitation pertains to the attributes of the tweets under examination. For the scope of this project, only 4 attributes are extracted and taken into consideration as mentioned in Section 2.3. However, other attributes such media attachments, language and promoted content on a tweet could also have potential influence in the way users interact with a tweet. Thus, it can affect the polarisation caused in the comments. The relationship between these attributes and users’ comments is not apparent and largely unknown.

One of the major challenges faced during data collection and scope definition were the presence of large amounts of hashtag spamming and bots on Twitter. This prevented the use of Twitter
APIs to fetch data using hashtag filtering as the tweets using the #RoeVsWade hashtag might not carry content relevant to our study.

Another one of the challenges faced while developing the web scraper is the rate-limiting imposed by Twitter on its API usage. To get the comments thread for a tweet, the conversation_id field is used in the GET method. Using Twitter API v2, this field is used to retrieve and reconstruct an entire conversation thread [40]. However, Twitter only allows up to 300 requests per application [41]. Therefore, to prevent abusing these rate limits, the data is stored locally in CSV format.

The initial plan for the data collection phase was to integrate the application with a database to overcome the Twitter API limits. The proposed plan was to use a PostgreSQL database with Python and store the extracted information in it. However, due to some knowledge and time constraints, this idea was discarded. Instead, a simple way to store the data – as CSV- was adopted.

5.2 Sentiment Analysis Techniques

In this project, we perform sentiment analysis through polarity and subjectivity detection, as well as text clustering. These techniques are important to determine for confirmation bias, as the approach uses polarity detection to identify potentially biased users, while also computing the likelihood ratio for confirmation bias. As we have observed in the above case studies, the overall performance in polarity and subjectivity detection methods are good as they were able to estimate the polarity and subjectivity based on some key words. However, as we have also seen in the above case studies, the polarity detection does not consider context and human emotions and hence, it can incorrectly estimate the polarity based on the true meaning of the comment, despite including terms that constitute a certain sentiment.

Secondly, we also observe that there exists comments where both the polarity and subjectivity detection models classify them in different classes. For instance, a comment can be classified as subjective by one method, and objective by another method. The inconsistencies in results for both techniques are not unexpected, since the methods employed in both techniques are different in its technical implementation. For instance, VADER and TextBlob are lexicon-based approaches while the BERT model method is a supervised learning model. Hence, the inconsistencies in the results from both techniques is a limitation of this application. In this situation, the comments would eventually be labelled as “Unknown” and cannot be used further in the analysis as “Unknown” comments are not identified for bias.

BERT sentiment analysis model has certain limitations when using for computation on social media data. Training and running the model requires a lot of computational resources and memory, which is time consuming for large datasets and inefficient for real-time applications [42]. BERT model may also face challenges with overfitting and degradation when fine-tuning on small or noisy datasets, impacting its performance. Additionally, BERT may struggle to capture nuances and subtleties of human emotions, such as irony, sarcasm, humour, or emotion
intensity. To overcome these limitations, additional features and techniques may be necessary to improve the performance of the BERT model [42].

Furthermore, we also observe that the text clustering technique did not perform well in distinguishing the comments in the different views. Instead, the text clustering algorithm looks for common words and assign the cluster to comments with similar words. This is because similar views are often not expressed in similar terms. The original assumption is that similar views are expressed in similar language to perform for text clustering. Hence, the text clustering method did not perform as expected too.

5.3 Identifying confirmation bias

Confirmation bias is identified and quantified based on the proposed criteria, namely the polarity of comments and the likelihood ratio. The likelihood ratio can allow us to determine the amount of confirmation bias as a whole within the conversation, as it shows the relative probability in one hypothesis over the other hypothesis with the same observation. However, this method is not extended to computing likelihood ratio for single user, as the approach does not include collecting the users’ retweets and replies of other conversations.

In order to extend this method of computing the likelihood ratio on the users, we would require the retweets, replies on other tweets, and information on likes of other tweets to compute the P(H), P(H’) and P(D) values. However, it is unfeasible to compute these values with the required information of likes, replies and retweets, because of the observation that a single user will not only react on a specific issue and can be on other topics as well. This is because a user can like, reply, and retweet on tweets that are of different topics to the comment that they have made in the specific tweet posted by a news provider. Additionally, the limits imposed by Twitter also makes it challenging to get more historic tweets of a specific user. As such in this project, only the comments within the conversation by a single user are used to calculate the likelihood ratio, while only analysing the users’ retweets, replies and likes as verification.

5.4 Verification of Results

To verify that the identification of confirmation bias is correct, the results verification can be done manually by checking the sentiment of the comments and to look at the behaviour of potentially biased users outside of the comment, by analysing their retweets, likes and replies. However, the verification process only looks at the filtered tweets, on the assumption that the users who are potentially biased only react to tweets that are related to the Roe vs Wade law in the United States. This assumption limits the scope of our analysis and may distort the results because some of these users can also react to other tweets from a range of topics which interest them as well as the views that are aligned with them. These tweets can also potentially allow us to discover more patterns on the users’ confirmation bias which can also be considered for analysis.
6. Future Works

This Chapter highlights a few ways in which this model can be extended and brought into work with Twitter in real-time.

6.1 Nudge Notifications

The sentiment analysis model can be integrated into a web browser extension or a mobile app that monitors the sentiment of tweets in real-time as users scroll through their Twitter feeds. When a tweet is identified as having a high likelihood of confirmation bias based on its sentiment, the system can generate a nudge notification, alerting the user about the potential presence of confirmation bias in the tweet. The notification can contain a gentle reminder to critically evaluate the information presented in the tweet and consider alternative perspectives.

6.2 Visual Cues

Another approach is to incorporate visual cues, such as color-coded indicators or icons, within the Twitter interface to highlight tweets that exhibit confirmation bias. For example, tweets with a positive sentiment that are likely to confirm users' existing beliefs could be marked in red, while tweets with a negative sentiment that challenge users' beliefs could be marked in green. This visual cue can serve as a subtle reminder to users to be mindful of the presence of confirmation bias in the content they encounter on Twitter.

6.3 Educational Pop-ups

The sentiment analysis model can also trigger educational pop-ups or tooltips that provide brief explanations about confirmation bias and its impact on decision-making. These pop-ups can appear when users click on tweets flagged by the sentiment analysis model as exhibiting confirmation bias, and can provide users with information on how confirmation bias can affect their perception of information, and encourage them to seek diverse sources of information to form a well-rounded opinion.

6.4 Personalised Recommendations

The results from the sentiment analysis model can be used to personalize users' Twitter experience by providing recommendations for diverse content that challenges their existing beliefs. By analysing the sentiment of the tweets users engage with and identifying potential confirmation bias, the system can recommend tweets with contrasting perspectives to help users broaden their understanding and reduce the impact of confirmation bias on their information consumption habits.

These are just a few potential ideas for extending the project to provide real-time nudges to Twitter users and raise awareness about confirmation bias. Further research and development needs to be carried out to refine and implement these ideas, especially by enhancing the text
clustering model to perform better with social media data and also by taking into consideration user feedback, ethical considerations, and technical feasibility.
7. Conclusion

In conclusion, our study aimed to analyse the presence of confirmation bias in social media comments on news articles or commentaries posted by news providers. We proposed an application that utilized various approaches, including data collection, pre-processing, and sentiment analysis, to identify and quantify confirmation bias. While we encountered strengths and limitations in implementing these methods, we also proposed alternative solutions to address the challenges.

Our application was tested on real-world social media data from PHEME, a collected dataset, as well as data collected from Twitter using Twitter APIs. The case studies provided valuable insights into the existence of confirmation bias in social media comments. We observed that some users expressed highly polarized sentiments, either positive or negative, indicating the presence of confirmation bias. Therefore, our developed model serves as a viable approach to quantify confirmation bias in social media. However, we also acknowledge that the scalability of our approach may be questionable, as discussed in Section 5. Further research and refinement of our methods may be needed to ensure robustness and scalability in different contexts.

Overall, our study contributes to the understanding of confirmation bias in social media discussions and provides insights into its existence. It also highlights the importance of considering alternative methods and addressing limitations in future research. Our findings have implications for social media users, news providers, and policymakers in promoting informed and unbiased discussions on social media platforms.
References


Appendix A: Case Study 1

The data obtained from PHEME is converted into a dictionary form with the necessary attributes:

User ID: 552832063614730824 Time: Fri Jan 10 22:03:30 +0000 2014
In reply to: 5528375611968656
@richardjgodwin Entirely accurate depiction of pasty baby britshits

User ID: 552842650099096969 Time: Sun Mar 01 20:42:33 +0000 2009
In reply to: 5528375611968656
@richardjgodwin rather worrying that was understood to be the English, even in caricature.

In reply to: 5528375611968656
@richardjgodwin @OllieHolt22 love the tattoo on the arm

User ID: 552845426278142464 Time: Fri May 18 13:37:34 +0000 2012
In reply to: 5528375611968656
@richardjgodwin @OllieHolt22 it called satire

User ID: 55284266474404352 Time: Sun May 22 12:19:34 +0000 2011
In reply to: 5528375611968656
@richardjgodwin @OllieHolt22 I know that woman! God bless Charlie Hebdo: keep taking the piss!

In reply to: 5528375611968656
@richardjgodwin @MichaelPDeacon harsh but fair

In reply to: 5528375611968656
@richardjgodwin @MichaelPDeacon As an Englishman I find this extremely insulting...but bloody funny! Ha ha ha ha

User ID: 55285233054002818 Time: Tue Oct 29 14:38:29 +0000 2013
In reply to: 5528375611968656
@richardjgodwin @BekkyC_RIP. Get ALL Islamist out of Civilization. It's either US or THEM.

The following table shows a subset of the dataset for the Charlie Hebdo shooting:

<table>
<thead>
<tr>
<th>id</th>
<th>timestamp</th>
<th>reply_to</th>
<th>comment</th>
<th>url</th>
<th>link_title</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>552832063614730824</td>
<td>Fri Jan 10 22:03:30 +0000 2014</td>
<td>@richardjgodwin Entirely accurate depiction of...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>552842650099096969</td>
<td>Sun Mar 01 20:42:33 +0000 2009</td>
<td>@richardjgodwin rather worrying that was under...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>552848414670034432</td>
<td>Mon Mar 09 15:28:09 +0000 2009</td>
<td>@richardjgodwin @OllieHolt22 love the tattoo o...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>552845426278142464</td>
<td>Fri May 18 13:37:34 +0000 2012</td>
<td>@richardjgodwin @OllieHolt22 it called satire</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>55284266474404352</td>
<td>Sun May 22 12:19:34 +0000 2011</td>
<td>@richardjgodwin @OllieHolt22 I know that woman...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>552839497160077112</td>
<td>Sat Oct 10 14:23:09 +0000 2009</td>
<td>@richardjgodwin @MichaelPDeacon harsh but fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>552840429904262528</td>
<td>Thu Jul 07 13:42:40 +0000 2011</td>
<td>@richardjgodwin @MichaelPDeacon As an Englishman...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>55285233054002818</td>
<td>Tue Oct 29 14:38:29 +0000 2013</td>
<td>@richardjgodwin @BekkyC_RIP. Get ALL Islamist...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following data tree represents the tree structure of the comments given a root comment:

```
5003832348810574849
  └── 50038304829269676832
     └── 5003836447390220289
        └── 5003836898864994817
            └── 5003836375373212242
                └── 500383761398188032
                    └── 5003837321067941888
                        └── 500383745345337856
                            └── 5003837493134804736
                                └── 5003837596411420973
                                    └── 5003837685559796802
                                        └── 500383969764275969
                                            └── 5003831888343269276
                                                └── 5003833796639096832
                                                    └── 500383859628413953
                                                        └── 500383866297599464
                                                            └── 5003839797393420400
```

The following table shows the results of sentiment analysis and confirmation bias:

43
| id  | comment textblob_polarity vader_compound_score vader_polarity modal_subjectivity textblob_subjectivity topic_cluster potential_bias |
|-----|--------------------------------------------------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 0   | @sharma_tbd @Rajapane Are you @budutt re tweeted that? What L... @dhuemee @CCIMR_brunn @Rajapane RSS goes too br... @dhuemee | 0.250000 0.00000 | NEU 0.198325 0.250000 0 0 |
| 1   | @Rajapane You @budutt re tweeted that? What L... | 0.300000 0.07722 | NEU 0.668814 0.000000 1 0 |
| 2   | @dhuemee @CCIMR_brunn @Rajapane RSS goes too br... | -0.100000 -0.42150 | NEG 0.061233 0.112500 1 1 |
| 3   | @dhuemee @Rajapane RSS goes too br... | 0.250000 0.23822 | NEU 0.989809 1.000000 1 0 |
| 4   | @UnvTrasSols @budutt @Rajapane Challenge if you... | 0.000000 -0.29182 | NEG 0.015514 0.000000 1 0 |
| 5   | @Rajapane Deleted not featured @budutt | 0.000000 0.00000 | NEU 0.154951 0.000000 1 0 |
| 6   | @dhuemee @Rajapane That's the difference between... | -0.350000 -0.94230 | NEG 0.996856 1.000000 0 1 |
| 7   | @rahulnriohan St. taking news should publish | 0.000000 -0.07722 | NEU 0.554125 0.000000 1 0 |
| 8   | "100 lashes if you don't die laughing" Out L... @parikshati @budutt @rajapane I cant comment | 0.500000 -0.17799 | NEU 0.738501 0.000000 1 0 |
| 9   | @rahulnriohan St. taking news should publish | 0.000000 0.00000 | NEU 0.871867 0.000000 0 0 |

...
<table>
<thead>
<tr>
<th>tweet_id</th>
<th>user1</th>
<th>user2</th>
<th>sentiment</th>
<th>confidence</th>
<th>score</th>
<th>retweets</th>
<th>replies</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>dhume @CEMB, flun</td>
<td>Worsome, any idea why no one is trend...</td>
<td>NEU</td>
<td>0.995062</td>
<td>0.000000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>Rajapre   @Ashish 2014</td>
<td>4030000</td>
<td>0.5106</td>
<td>POS</td>
<td>0.480306</td>
<td>0.800000</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>Rajapre   @BDUTT This is price the press pact...</td>
<td>0.200000</td>
<td>-0.6124</td>
<td>NEG</td>
<td>0.000123</td>
<td>0.000000</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>Rajapre   @BDUTT what does this mean? Don't u...</td>
<td>0.158250</td>
<td>0.0000</td>
<td>NEU</td>
<td>0.05362</td>
<td>0.343750</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>Rajapre   be careful Dhume now. Follow...</td>
<td>0.100000</td>
<td>-0.6249</td>
<td>POS</td>
<td>0.000200</td>
<td>1.000000</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>Rajapre   Hope you understand...</td>
<td>0.000000</td>
<td>0.4404</td>
<td>POS</td>
<td>0.407344</td>
<td>0.000000</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>Rajapre   The 2011 issue of Charlie Hebdo...</td>
<td>0.000000</td>
<td>0.0000</td>
<td>NEU</td>
<td>0.001917</td>
<td>0.000000</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>Rajapre   u might get sh...</td>
<td>0.000000</td>
<td>-0.4019</td>
<td>NEG</td>
<td>0.968882</td>
<td>0.000000</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>Rajapre   only true...c...</td>
<td>0.000000</td>
<td>-0.6039</td>
<td>NEG</td>
<td>0.070057</td>
<td>1.000000</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix B: Case Study 2

The data obtained from Twitter is converted into a dictionary form with the necessary attributes:

User ID: 1505183608562988860 Time: 2023-04-13T03:45:01.000Z
In reply to: 30354991
@KamalaHarris @TheDemocrats https://t.co/OCv1Vc6jFu

User ID: 1643706647346328064 Time: 2023-04-13T01:52:37.000Z
In reply to: 30354991
@KamalaHarris Pull your head out of your ass

In reply to: 30354991
@KamalaHarris It’s time for you to leave office

User ID: 1628534482444918593 Time: 2023-04-12T23:12:11.000Z
In reply to: 30354991
@KamalaHarris It’s been time. But later it better than never.

User ID: 1574085941414010900 Time: 2023-04-11T01:38:08.000Z
In reply to: 30354991
@KamalaHarris No that’s just stupidity

In reply to: 30354991
@KamalaHarris It time for you to start doing your job, do you know what that is?

User ID: 182632692 Time: 2023-04-09T02:14:51.000Z
In reply to: 30354991
@KamalaHarris Are you this dumb? It is a state issue.

User ID: 1574085941414010900 Time: 2023-04-09T00:12:42.000Z
In reply to: 30354991
@KamalaHarris Go to hell you devil woman!!!

User ID: 158629017647014721 Time: 2023-04-08T13:04:37.000Z
In reply to: 30354991
@KamalaHarris The Supreme Court says otherwise!

The following table shows a subset of the dataset for the Kamala Harris tweet:

<table>
<thead>
<tr>
<th>id</th>
<th>timestamp</th>
<th>reply_to</th>
<th>comment</th>
<th>social_media</th>
<th>conversation_id</th>
<th>head_id</th>
<th>user_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2023-04-13T20:07:28.000Z</td>
<td>30354991</td>
<td>@KamalaHarris @TheDemocrats <a href="https://t.co/yM7L">https://t.co/yM7L</a>...</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1595138669506280160</td>
</tr>
<tr>
<td>1</td>
<td>2023-04-13T15:26:25.000Z</td>
<td>30354991</td>
<td>@KamalaHarris <a href="https://t.co/loVegYDQjPa">https://t.co/loVegYDQjPa</a></td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>13310243556421893665</td>
</tr>
<tr>
<td>2</td>
<td>2023-04-13T03:45:01.000Z</td>
<td>30354991</td>
<td>@KamalaHarris @TheDemocrats <a href="https://t.co/OCv1V">https://t.co/OCv1V</a>...</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1595138669506280160</td>
</tr>
<tr>
<td>3</td>
<td>2023-04-13T01:52:37.000Z</td>
<td>30354991</td>
<td>@KamalaHarris Pull your head out of your ass</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>16437926674366328664</td>
</tr>
<tr>
<td>4</td>
<td>2023-04-12T23:22:56.000Z</td>
<td>30354991</td>
<td>@KamalaHarris It's time for you to leave office</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1331480587630587906</td>
</tr>
<tr>
<td>5</td>
<td>2023-04-12T23:12:11.000Z</td>
<td>30354991</td>
<td>@KamalaHarris It's been time. But later it bet...</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1628534482444918593</td>
</tr>
<tr>
<td>6</td>
<td>2023-04-11T01:38:08.000Z</td>
<td>30354991</td>
<td>@KamalaHarris No that's just stupidity</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1574040594141401090</td>
</tr>
<tr>
<td>7</td>
<td>2023-04-09T12:46:26.000Z</td>
<td>30354991</td>
<td>@KamalaHarris It time for you to start doing s...</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1521320092905872384</td>
</tr>
<tr>
<td>8</td>
<td>2023-04-09T01:52:37.000Z</td>
<td>30354991</td>
<td>@KamalaHarris Are you this dumb? It's a stat...</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1574040594141401090</td>
</tr>
<tr>
<td>9</td>
<td>2023-04-09T20:12:42.000Z</td>
<td>30354991</td>
<td>@KamalaHarris Go to hell you devil woman!!!</td>
<td>Twitter</td>
<td>1643638127804796931</td>
<td>30354991</td>
<td>1674040594141401090</td>
</tr>
</tbody>
</table>
The following table shows the results of sentiment analysis and confirmation bias:

<table>
<thead>
<tr>
<th>id</th>
<th>id</th>
<th>user</th>
<th>comment</th>
<th>textblob_polarity</th>
<th>vader_compound_score</th>
<th>vader_polarity</th>
<th>model_subjectivity</th>
<th>textblob_subjectivity</th>
<th>topic_cluster</th>
<th>potential_bias</th>
</tr>
</thead>
</table>
Appendix C: Case Study 3

The data obtained from Twitter is converted into a dictionary form with the necessary attributes:

The following table shows a subset of the dataset for the Kamala Harris tweet:

<table>
<thead>
<tr>
<th>id</th>
<th>timestamp</th>
<th>reply_to</th>
<th>comment</th>
<th>social_media</th>
<th>conversation_id</th>
<th>head_id</th>
<th>user_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2023-04-10T22:29:09.000Z</td>
<td>3047925973</td>
<td>@OrianWeissman Using religious beliefs to justify forcing women to give birth to babies w fatal abnormalities violates the 1st Amendment by requiring women...</td>
<td>Twitter</td>
<td>3047925973</td>
<td>829063288566714377</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2023-04-08T01:53:55.000Z</td>
<td>3047925973</td>
<td>@OrianWeissman We shouldn't have to share our most private &amp; heartbreaking stories to get those with power in this country to respect OUR existence, OUR...</td>
<td>Twitter</td>
<td>3047925973</td>
<td>156931927428557135</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2023-04-08T08:38:32.000Z</td>
<td>3047925973</td>
<td>@OrianWeissman Sounds like they had a crap OB who didn’t stick their neck out to protect the mother and help her find alternate options. But no surprise, th...</td>
<td>Twitter</td>
<td>3047925973</td>
<td>213389971</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2023-04-07T23:49:19.000Z</td>
<td>3047925973</td>
<td>@OrianWeissman I should have had that choice - that right over my own body and over my daughter’s body to be able to tell my daughter, ‘It is time for you...</td>
<td>Twitter</td>
<td>3047925973</td>
<td>1117968494</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2023-04-07T23:35:29.000Z</td>
<td>3047925973</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>Twitter</td>
<td>3047925973</td>
<td>154711971584858570</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2023-04-07T23:35:14.000Z</td>
<td>3047925973</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>Twitter</td>
<td>3047925973</td>
<td>154711971584858570</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2023-04-07T23:35:14.000Z</td>
<td>3047925973</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>Twitter</td>
<td>3047925973</td>
<td>154711971584858570</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2023-04-07T23:35:14.000Z</td>
<td>3047925973</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>Twitter</td>
<td>3047925973</td>
<td>154711971584858570</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2023-04-07T23:35:14.000Z</td>
<td>3047925973</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>Twitter</td>
<td>3047925973</td>
<td>154711971584858570</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2023-04-07T23:35:14.000Z</td>
<td>3047925973</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>Twitter</td>
<td>3047925973</td>
<td>154711971584858570</td>
<td></td>
</tr>
</tbody>
</table>

The following data tree represents the tree structure of the comments given a root comment:
The following table shows the results of sentiment analysis and confirmation bias:

<table>
<thead>
<tr>
<th>id</th>
<th>comment</th>
<th>textblob_polarity</th>
<th>vader_compound_score</th>
<th>vader_polarity</th>
<th>model_subjectivity</th>
<th>textblob_subjectivity</th>
<th>topic_cluster</th>
<th>potential_bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>@DranWeissman Using religious beliefs to just...</td>
<td>0.000000</td>
<td>-0.8658</td>
<td>NEG</td>
<td>0.002914</td>
<td>0.250000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>@DranWeissman We shouldn’t have to share our...</td>
<td>0.250000</td>
<td>0.0094</td>
<td>NEU</td>
<td>0.001405</td>
<td>0.437500</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>@DranWeissman Sounds like they had a crap OB...</td>
<td>-0.400000</td>
<td>0.0974</td>
<td>NEU</td>
<td>0.139000</td>
<td>0.400000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>@DranWeissman ❤️</td>
<td>0.000000</td>
<td>0.0000</td>
<td>NEU</td>
<td>0.567332</td>
<td>0.000000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>@DranWeissman &quot;I should have had that choice...</td>
<td>0.461905</td>
<td>0.0000</td>
<td>NEU</td>
<td>0.002586</td>
<td>0.720238</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>@LisaMikolajczyk That’s the brand</td>
<td>0.000000</td>
<td>0.0000</td>
<td>NEU</td>
<td>0.026527</td>
<td>0.000000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>@DranWeissman Inhumane 💔 heartless to thak...</td>
<td>-0.900000</td>
<td>-0.8225</td>
<td>NEG</td>
<td>0.299191</td>
<td>0.300000</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>@DranWeissman This is heartbreaking, No one is...</td>
<td>-1.000000</td>
<td>-0.8769</td>
<td>NEG</td>
<td>0.004558</td>
<td>1.000000</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>@DranWeissman Tyranny of the...</td>
<td>0.350000</td>
<td>0.0000</td>
<td>NEU</td>
<td>0.001500</td>
<td>0.550000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>@DranWeissman Horrendous.</td>
<td>0.000000</td>
<td>-0.5869</td>
<td>NEG</td>
<td>0.996613</td>
<td>0.000000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>