TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Simon Fox Kuuse

Studying bias in Twitter (X) Community Notes

Bachelor Thesis (9 ECTS)

Supervisor(s): Uku Kangur MSc,
Roshni Chakraborty PhD

Tartu 2024

Kallutatuse uurimine Twitteri (X) Community Notedes

Lühikokkuvõte:

Faktikontrolle vajava sisu pidevast kasvust tingituna on pakutud välja uusi ühisloomel põhinevaid lahendusi. Selle töö eesmärk on ühes sellises lahenduses, nimelt Twitteri poolt loodud Community Notedes, potentsiaalse kallutatuse tuvastamine. Töö läbiviimiseks koguti kaks andmestikku, mis koosnevad Community Notedest ja allikate kallutatusest. Töös kasutatakse keelelist analüüsi, meelestatuse analüüsi, ajalist analüüsi ja võtmesõnade eraldamist. Tulemused näitavad korrelatsiooni kallutatuse ning meelestatuse vahel ja samuti kallutatuse ning päriselu sündmustele reageerimise vahel, mis võib viidata potentsiaalsele kallutatusele Community Notedes.

Võtmesõnad:

Sotsiaalmeedia, Meelestatuse Analüüs, Ajaline Analüüs, Võtmesõnade Eraldamine, faktikontroll, Community Notes, Kallutatuse Tuvastus

CERCS: P170 Arvutiteadus, arvutusmeedodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Studying bias in Twitter (X) Community Notes

Abstract:

With the frequency of content in need of fact-checking constantly rising, new approaches in the form of crowdsourced fact-checking are being tested. In this thesis, we aim to identify potential bias in one of Twitter's approaches, called Community Notes. Two datasets, comprising Community Notes and bias ratings of sources, are collected. We utilise lexical features, sentiment analysis, temporal analysis and keyphrase extraction. Our study shows a correlation between the sentiment and reaction to real-world events and the bias of a Community Note, suggesting that a potential bias is present.

Keywords:

Social Media, Sentiment Analysis, Temporal Analysis, Keyphrase Extraction, fact-checking, Community Notes, Bias Detection

CERCS: P170 Computer science, numerical analysis, systems, control

Table of Contents

1. Introduction

Bias is defined as the presence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics, which could lead to different outcomes for different individuals or groups [1]. For example, bias could be a strong feeling in favour or against a group of people or a side of an argument which is not based on fair judgement [2]. It can result from one side of the argument receiving more coverage or from an author supporting one side of the story or fabricating information [3]. Bias can be of different forms, such as demographic bias, i.e., underrepresentation of certain demographic groups or where the trend promoters do not represent the demographics of the media site's overall population, etc [4]. Similarly, media coverage or bias in news media, i.e., slanted representation of news events, can severely impact public perception, affect elections, and cause segregation and polarisation in society [3], [5]. Several existing research works have proposed automated mechanisms to identify and mitigate different types of media bias, including political leaning based on news articles [6], [7].

Recently, information-seeking behaviour has drastically changed from offline news media agencies to online modes of information, such as social networks and online news media agencies [8]. While social networks provide continuous access to vast volumes of information, different aspects of user opinions and stance, they also have increased the challenges manifold. For example, it can adversely impact public perception and user choices, sway national elections, spread misinformation and hate speech, and cause segregation of users into echo chambers [9], [10], [11]. For example, selective information exposure on the basis of political leaning has led to echo chambers on Twitter [12]. Additionally, bias on Twitter might not always be intentional, and studies show that over 80% of misinformation sharing can be attributed to inattention and confusion [13]. Therefore, it is essential to fact-check information on Twitter to reduce the effects of bias and misinformation. However, there are several challenges in automatic bias detection on Twitter due to the unprecedented volumes of information generated continuously. One of these challenges is ensuring unbiased training data when training machine learning models [11], [14]. Information such as the demographics of the posting user and their location can lead to better approaches but are not available due to privacy concerns [15]. The vast amount of information on social media, in comparison to traditional news media, makes it infeasible for professional fact-checkers to ensure the accuracy of the information shared. These fact-checks, performed by outside sources also face the issue of visibility and accessibility,

with studies showing fact-checking stories as having little to no influence over online news media agenda [16], [17].

Community Notes (previously known as “Birdwatch”) is a crowdsourced fact-checking solution that is developed by Twitter. It lets users add context to tweets or warn others of potential misinformation [18]. A Community Note can be up to 2500 words long and its title is limited to 100 characters [19]. Starting from 18.10.2023, all Community Notes that add context to a tweet are also required to link a source [20]. Community Notes are unique in that they are only publicly shown if enough people of differing opinions mark a Community Note as helpful [21]. The general idea of this is the wisdom of the crowd, so a Community Note is checked by a distributed group of users before being shown to all users to ensure the quality of the Community Notes. To achieve this, Twitter requires all Community Notes to be marked as helpful by a group of people who have disagreed with each other in the past to be shown [21]. These people need to be registered as Community Notes contributors and Community Notes with not enough votes on their helpfulness will only be shown to them, if enough people vote the Community Note as not helpful, it is hidden [22]. This approach allows context that people with different points of view find helpful [23]. Previous works have analysed the effectiveness of Community Notes in stopping the spread of fake news, with some works showing positive results [24].

To ensure the fact-checking done in the Community Notes is helpful and accurate, detecting bias in them is pivotal. In this thesis, we are studying the bias in Community Notes. For this, a public dataset containing all Community Notes posted between 28.01.2021 and 11.11.2023 is used to evaluate the bias in Community Notes. We assign each Community Note a label, i.e., Left-Leaning, Centre or Right-Leaning, based on the sources linked in the Community Note. This is the base of the work and is used together with all methodologies to find differences in the usage of Community Notes by bias group.

The first of these methods is readability analysis, which we use to find correlations between the readability of Community Notes and their bias. Sentiment analysis is utilised at word-, sentence-, and note-level to determine how positive and negative Community Notes with differing bias labels are. Temporal analysis is performed to analyse changes in Community Notes over time by looking at what prompted changes in the Community Notes posted by different bias groups. We observe that there is a higher occurrence of negative Community Notes in the left-leaning group compared to the centre group. Analysing the keywords of two weeks with highly polarising word usage indicates that left-leaning Community Notes are

more connected to real-world events, especially cases of gun violence, compared to right-leaning Community Notes. These findings show promising results in identifying bias in Community Notes, with clear differences between bias groups being observed using the aforementioned methods.

This thesis is split into four parts. The related works section covers the previous research into bias and Community Notes. The data section explains how the base Community Notes are collected and preprocessed along with how the bias of sources is determined. The methodology section explains each method, its use, and why it is used. In the last section, the results of the work are shown, along with relevant discussions.

2. Related works

The related works chapter highlights previous research in the area, limitations of the previous works and motivations for the work. We explain the concept of bias, its occurrence on social media platforms and what has been researched regarding Community Notes.

2.1. Bias

With the widespread usage of artificial intelligence and automated machine learning-based models in real-life systems, understanding the inherent bias is highly crucial in recent times irrespective of the application. There can be different forms of bias, such as demographic bias, representation bias, information exposure bias, media coverage bias, etc [1]. For example, demographic bias highlights the difference on the basis of gender, race, ethnicity, parental educational background, etc [1], [25]. Similarly, with the prevalence of bias in news coverage and representation, politically aligned news article reporting and the spread of propaganda through news articles, media bias detection and mitigation has become highly relevant [9], [26], [27]. Several research works highlight the possible adverse effects of media bias on society and propose computational approaches to identify bias in news articles, such as identifying bias-inducing words in online news media [26], [27], assessing the ideological stances [28], political leaning based bias detection [26], [29], etc. Therefore, these works highlight the requirement of identifying bias and the different aspects to be prioritised during fact-checking [30].

2.2. Bias in Social Media Platforms

With the huge popularity of online social networks, such as Twitter, Facebook, Reddit, etc., news media agencies have also shifted to Twitter to publish their news articles [31]. Additionally, a huge fraction of users receive their daily news from social networks [32]. It has been observed that users' actions on Twitter show their real-world political behaviour [33]. Furthermore, users have access to diverse sources of information on Twitter through indirect media exposure from friends and retweeted posts, which has been shown to increase the political diversity of information that users are exposed to [34]. However, there are several severe adverse impacts from news consumption from Twitter, such as the advantage of the most influential users on Twitter who hold significant influence on information spread [35] and the high spread of misinformation. Findings, such as the top 1% of false news on Twitter reaching significantly more people than the top 1% of accurate news, highlight the

need for a fast fact-checking solution [36]. Subsequently, it has been observed that biassed information exposure through selective information exposure and confirmation bias leads to the formation of echo chambers and segregation of users [12]. It has been observed that users tend to share and engage with content that corresponds with a specific narrative [37]. For example, users that are exposed to some of these communities, like ones consisting of conservative users, have been shown to be exposed to more low-credibility content [38]. Subsequently, the spread of propaganda remains a concern on Twitter, such as, only 8-15% of propaganda was removed from Twitter effectively with respect to the Russian invasion of Ukraine event [39]. Although the relative ease of data collection from Twitter in comparison to traditional news media sources aids in identifying bias in news article reporting through manual verification and understanding of different user perspectives [31], the vast volume of information generated makes manual verification severely challenging. Automated solutions to this moderation have been proposed, but issues such as missing context and not ensuring the legibility of available information lead to inaccurate misinformation detection [40], [41], [42]. Therefore, identifying alternative solutions which aid in bias mitigation and fact-checking on Twitter is needed along with an understanding of the inherent biases in the proposed solutions.

2.3. Community Notes

There have only been a few works related to Community Notes thus far, which have primarily been simple exploratory and comparative studies with the data.

Even these few works have come to differing conclusions, with one finding that there has not been a reduction in the engagement of false news on Twitter [18]. Findings of another paper suggest a 50% reduction in retweets and an 80% increase in deletes of tweets with a Community Note attached [24]. Overall findings agree that the slow speed of Community Notes appearing on tweets might hinder their effectiveness. The targets of Community Notes have also been explored, with the findings showing a higher occurrence of Community Notes on larger accounts. The findings also showed that Community Notes are usually used to refute and not agree with an argument [43]. However, no works have analysed how biased the fact-checks within the Community Notes are. Therefore, in this Thesis, this is explored and studied exhaustively.

3. Dataset Details

This section describes the data collection process from Twitter and Community Notes. Twitter is a platform where users can share their opinions in the form of tweets. A tweet is a short text of up to 280 characters that can also include a picture, a link or a survey. Users also have the ability to like, retweet, share or bookmark tweets [44]. The Community Notes dataset is discussed next in detail.

3.1 Community Notes

Source bias data is collected by downloading all Community Notes that were posted between January 23, 2021 and November 11, 2023. This data is made publicly accessible to download on the Twitter Community Notes website[1] and contains 323 382 Community Notes. The available dataset consists of 4 different files, namely:

- Notes - Contains general info about the Community Note, e.g. the creator, creation time, contents of the Community Note and why the Community Note was added.
- Ratings - Contains the ratings of the note, added by other Twitter users.
- Note Status History - Contains information about when and what kind of ratings the note has gotten.
- User Enrollment - Contains information about the person who submitted the Community Note, e.g., how well others have rated their notes collectively.

For this work, only the data in the Notes file was used as it contains the text of each Community Note and, along with it the sources, which were used to determine the bias [45]. This data contains no personal information and the results are presented in an aggregated way so no person's data can be identified.

[1] https://communitynotes.twitter.com/guide/en/under-the-hood/download-data

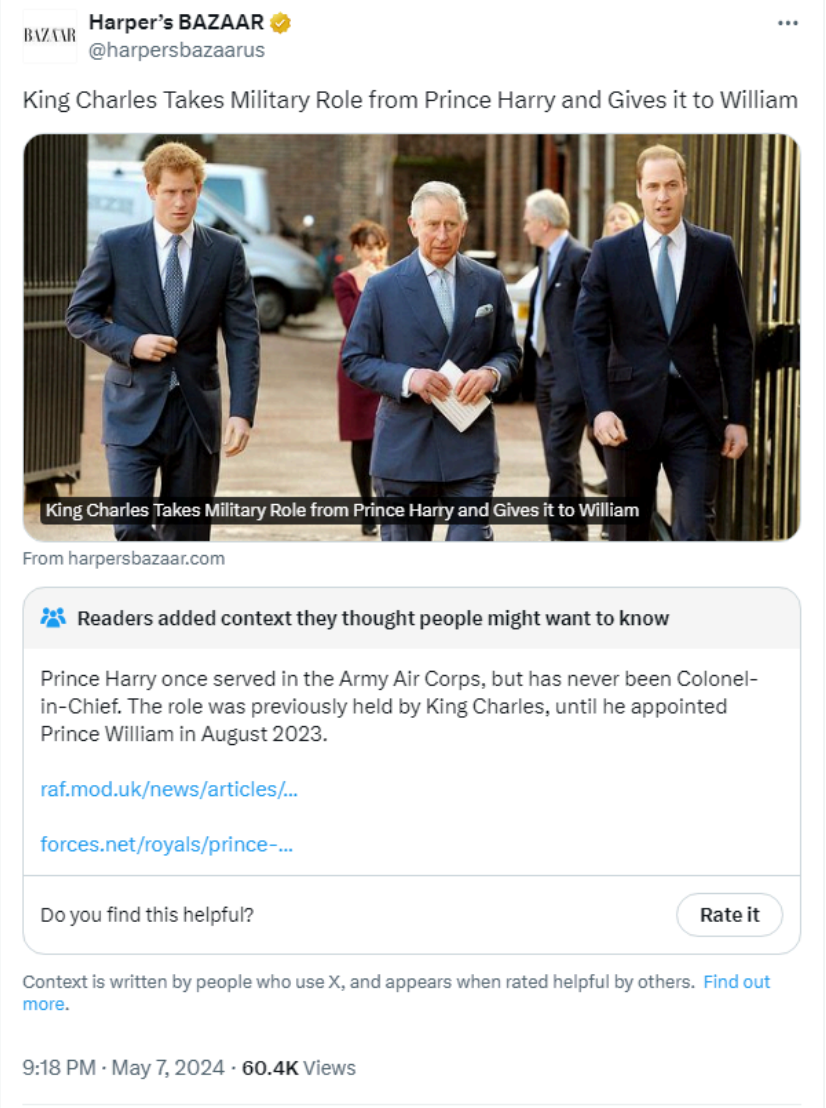

Figure 1. Screenshot of a tweet and the Community Note added to it [46].

3.2 Dataset Preprocessing Details

The downloaded Community Notes dataset was read into Python[2] from TSV files and then converted to a dataframe using the Pandas[3] library. Out of the original columns, two were kept that included the unique id of every Community Note and the contents of the Community Note with references to the sources. The original dataset contained a large amount of noise in the form of Community Notes that were not in English and Community Notes that did not have any relevant sources linked. Each Community Note text was preprocessed, after which a new column with the preprocessed data was added to the dataset. To identify and remove Community Notes that were not in English, the langdetect[4] of the

[2] https://www.python.org/
[3] https://pandas.pydata.org/
[4] https://github.com/Mimino666/langdetect

library was used. Words belonging to the NLTK[5] *stopwords* set were removed. The set consists of words that occur very frequently and don't add much info about the text being worked on, for example, each, about, such, and the [47]. To process words with the same stem as one, the NLTK *EnglishStemmer* algorithm was used. In each text, stems that occurred more than once were detected and replaced with their stem. After preprocessing, the final dataset comprises 35 536 Community Notes out of the original 323 382.

Relevant Domain Extraction: We perform several pre-processing steps before bias-level annotation. This needs to be done as links to the same domain have matching biases and can be grouped as one, we then compare the groups to find the most used sources. To group links that represent the same domain, we first shorten all URLs to their hostnames, e.g. https://www.example.com/article/123 to example.com. Then a script is used to expand three types of links: link shorteners, e.g. *bit.ly*, social media redirects, e.g. *goo.gl*, and web archives, e.g. *archive.is* to find their original link. After this initial pre-processing, we group URLs that share the same domain. For example:

- Versions of the same website for different devices, e.g. *en.wikipedia.org* and *en.m.wikipedia.org*
- Short and long forms of the websites, e.g. *youtube.com* and *youtu.be*
- Same websites with different top-level domains, e.g. *bbc.com* and *bbc.co.uk*
- Subdomains of the same websites, e.g. *twitter.com* and *help.twitter.com*

From the groups obtained using these methods, we keep the 500 groups with the highest frequency of usage. These groups consist of 4064 URLs that occur in 306 576 Community Notes.

3.3 Bias Annotation

In this subsection, the bias label annotation procedure is discussed in which each Community Note is annotated as right-leaning, left-leaning or centre.

Bias Label Identification:

Each Community Note needs to be annotated with a bias label to analyse the Community Notes in relation to bias. As we assign the bias labels to Community Notes based on the bias of their references, a Community Note cannot be assigned to a bias group without any valid sources. To find the bias of these sources, we aggregate the bias labels from three media

[5] https://www.nltk.org/

monitoring websites: *mediabias-factcheck.com, allsides.com, and adfontes.com*. A Community Note is assigned the label on the basis of the majority of the labels assigned by the media monitoring websites. However, if no labels occur more often than others, the source is removed from the dataset. After bias-level annotation, the dataset comprises 306 576 Community Notes and 183 sources, which contain 991 URLs.

To label the Community Notes we first preprocess the data to remove all rows that do not refer to a valid source or contain both left- and right-leaning sources at the same time. Community Notes that have sources of differing bias labels can not be reliably placed in a group as they have elements of both left-leaning and right-leaning bias. We then assign each Community Note a bias label of left-leaning, centre or right-leaning, based on the average bias of its sources, based on the following principle.

- If the average bias score of a Community Note's sources is greater than 0.5, the Community Note is labelled as *left-leaning*.
- If the average bias score of a Community Note's sources is greater than or equal to -0.5, but less than or equal to 0.5, the Community Note is labelled as *centre*.
- If the average bias score of a Community Note's sources is less than -0.5, the Community Note is labelled as *right-leaning*.

To give the sources a score, the source bias data was read into Python from CSV files and was then converted to a dataframe using the Pandas library. Columns that included the URL of the source and the bias of the source were kept. The bias scores of sources were converted into numerical values for easier processing, and the principle can be seen in Table 1 with a visualisation of the number of sources by bias label in Figure 2:

Table 1. Conversion of bias labels into a numerical format

Original Label	New Label	Count
Left	2	14
Left-Center	1	66
Center and Pro-Science	0	73
Right-Center	-1	22

Right	-2	8

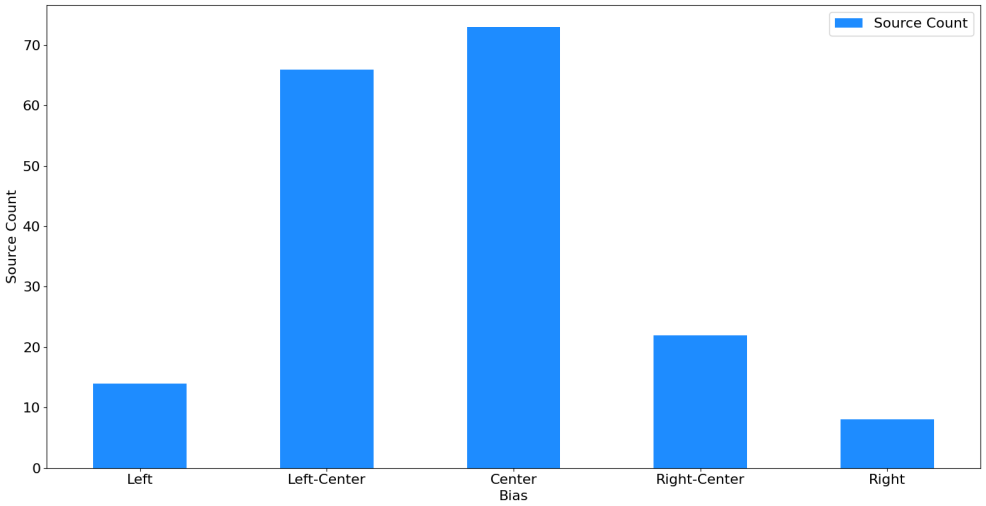

Figure 2. Visualisation of the number of sources by bias label

4. Methodology

In this thesis, we explore different types of methods to evaluate the bias of Community Notes. These methods include lexical features, topic modelling, sentiment analysis, and temporal analysis, which we will discuss in detail next. In this Section, we explain these methods in detail next, along with the basis for their usage. We additionally dive into the implementation of these features by highlighting the libraries that were used. We have added the input and the form of the output for each method to give a better understanding of their usage.

4.1 Lexical features

We use lexical features to better understand the context in which words are used, remove noise, and evaluate the style of writing used in the Community Notes concerning bias identification for Community Notes. In this thesis, we analyse lexical features using POS tagging and readability analysis, which we discuss in detail next. We remove stopwords from the Community Notes text as pre-processing (details explained in Subsection 3.2).

Adjectives: POS Tagging is the method of tagging words based on the part of the paragraph they are in and the context of the word usage. Examples of possible POS Tags, their meanings and corresponding words are shown in Table 2. In this work, we focus on adjectives as they have been shown to have a correlation with the subjectivity of the text [48]. To understand the usage of adjectives between Community Notes of different bias groups, we identify the average frequency of adjectives used in a Community Note for each group. Adjectives identified are also used for sentiment analysis and the results are available in subsection 5.3.

Table 2. Examples of possible POS Tags, their meanings and corresponding words.

JJR	adjective, comparative	bleaker, cuter
NNP	noun, proper, singular	Darryl, Shannon
RB	adverb	occasionally, prominently

Readability Analysis: The readability of a text is imperative in garnering a large audience. The reader's ability to understand the text depends on multiple factors, e.g. the reader's level of education, time spent by the reader, previous experience with the topic, and motivation [49]. Our purpose in observing the readability is to find a possible correlation between the difficulty of the text and its bias. We judge the readability of the text in 4 different ways:

- *Flesch Reading* - Equation 1, used for this score, is the following:

$$FRS=206.835-1.015\left(\frac{total\ words}{total\ sentences}\right)-84.6\left(\frac{total\ syllables}{total\ words}\right)$$

Equation 1. Flesch Reading score (FRS)

The theoretical lowest score is -3.4, and there is no upper limit. A lower score correlates to a higher difficulty of comprehension.

- *Dale Chall* - Equation 2, used for this score, is the following:

$$DCS=0.1579\left(\frac{difficult\ word}{words}\times 100\right)+0.0496\left(\frac{words}{sentences}\right).$$

Equation 2. Dale Chall score (DCS).

A word is considered difficult if a fourth-grade student can't reliably comprehend it. The theoretical lowest score is 0.0496, and there is no upper limit. A higher score correlates to a higher difficulty of comprehension.

- *Coleman Liau* - Equation 3, used for this score, is the following:

$$CLS=0.0588 \times L - 0.296 \times S-15.8$$

Equation 3. Coleman Liau score (CLS).

L is the average number of letters per 100 words, and S is the average number of sentences per 100 words. A higher score correlates to a higher difficulty of comprehension.

We utilise the *readability*[6] library to grade the readability of the text. This method assigns each Community Note a readability score and corresponding educational level. Our observations and results for readability analysis are in Subsection 5.2

[6] https://github.com/andreasvc/readability/

4.2 Sentiment analysis

Sentiment analysis is the method of modelling people's opinions, attitudes, and feelings, toward individuals, events, or topics. The purpose of the method is to find what sentiment a text represents and find the estimate of the text on a scale from negative to positive [50]. These findings are important to compare the type of writing between Community Notes with differing bias labels and the results can be observed in subsection 5.3. In this Thesis, we perform sentiment analysis using word-based sentiment, sentence-based sentiment and note-based sentiment, which we discuss in detail below.

Word-Based Sentiment Analysis: We consider only those words that are tagged as adjectives to display sentiment. We initially split the Community Note into individual words, consider only those words which are adjectives and finally, grade each adjective on a positive and negative scale, which are both in the 0-1 range on the basis of *sentiwordnet*[7] library. This will give us the frequency of positive, negative and neutral words in every Community Note.

Sentence-Based Sentiment Analysis: We approached sentiment analysis of sentences in 2 ways. First, we assign each sentence in a Community Note a label of positive, negative or neutral. We then count the occurrence of positive, negative and neutral sentences in a Community Note. This enables us to find the number of positive, negative and neutral sentences per community note. Based on this, we can find the total frequency of sentences with a given sentiment for a bias label, e.g. negative sentences in left-leaning Community Notes. For the second approach, we compare the number of sentences with a given sentiment in each Community Note. If there are more negative sentences than neutral or positive, the Community Note is marked as negative. If there are more positive sentences than neutral or negative, the Community Note is marked as positive. In all other cases, the Community Note is marked as neutral.

To assign each sentence a sentiment label, we use the VADER-Sentiment-Analysis tool. This is an open-sourced tool that is specifically made to analyse the sentiment in social media. For each sentence in a document, a compound score is returned, this score is constrained between -1 and 1 [51]. We adopt the standard approach of grading sentences based on this compound score in the following way:

- Compound score $\geq$ 0.05, the sentence is labelled as positive.

[7] https://github.com/aesuli/SentiWordNet

- Compound score < 0.05 and compound score > -0.05, the sentence is labelled as neutral.
- Compound score ≤ -0.05, the sentence is labelled as negative.

Note-Based Sentiment Analysis: To find the sentiment of the whole Community Note, a Hugging Face model, specifically the *cardiffnlp/twitter-roberta-base-sentiment-latest*[8], which is trained on ~124M tweets from January 2018 to December 2021. This model is trained on a format similar to Community Notes and displays a very good ability to categorise the Community Notes when manual checking is done on a sample of the dataset. The model also outputs a confidence score between 0-1 on how close a certain Community Note is to a bias, to further eliminate any outliers, only Community Notes with a confidence score greater than 0.7 are kept for analysis utilising these values.

4.3 Temporal analysis

Temporal Analysis: Temporal Analysis is the study of changes in data over time. These results are important to find how different bias groups react to events and how their behaviour changes over time. The first goal of temporal analysis is to find events or topics that prompt a higher usage of polarising words in bias groups. Additionally, we compare if the response to these events or topics depends on the Community Notes' bias. The second goal of utilising temporal analysis is to find long-term changes in the polarity of Community Notes of a given bias label. This is important for identifying trends and possible events that lead to these changes. For temporal analysis, the findings of sentiment analysis are used, which has been explained in subsection 3.2. The main data point for this analysis is the polarity of the Community Note, which is found using Equation 4.

$$P = \frac{PW+NW}{TW}$$

Equation 4. Used to find the polarity of a Community Note

P is the polarity of a Community Note, PW is the frequency of positive words in it, NW is the frequency of negative words in it and TW is the total frequency of words in it.

For the first goal, we find weeks in the data where the average polarity of Community Notes with a given bias label is higher than the mean. These periods are important for finding real-world events that lead to an increased usage of polarising words in Community Notes. We use the one-week-long period for two reasons. Firstly a shorter period has too much noise

[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

in the form of some groups having less than 5 Community Notes, which increases the deviation of the polarity. Secondly, a longer timeframe is avoided to better identify individual events that lead to an increase in polarity. For the second goal, a linear regression line is constructed based on the polarity of Community Notes.

4.4 Keyphrase Extraction

In natural language processing, topic models are generative models, which provide us with possible topics for a document and its odds of belonging to each topic. Topics signify the relations between words in a vocabulary and their occurrence in documents [52]. The purpose of topic modelling in this work is to identify the most prevalent underlying topics that are discussed in Community Notes. We analyse topic modelling through keyphrase extraction which we discuss in detail next.

Keyphrase Extraction: A keyphrase is a series of words that has high relevance to a corresponding document and therefore represents the whole document. A subset of all Community Notes that have the same bias label and are posted during a given time period is used as the document in this work and this will be referenced to as the corpus. We use keyphrase extraction to find keyphrases that can best represent the corpus and by doing that find out the most popular themes in these Community Notes. To extract keyphrases from our data the *Keybert*[9] library is utilised which identifies the keyphrases that are present in the Community Notes. For this, we consider the corpus, remove the URLs from the base text and then merge all of the Community Notes into a single string. We then run the model with this string as the input, along with this we also set the length of the keyphrases that are returned. The outputs of this model are in the form of keyphrase and score. The score is the relevance of the keyphrase in relation to the Community Notes and is constrained to range from 0 to 1.

Temporal Topic Analysis: We want to find the keywords that best represent all Community Notes posted during a week for both bias groups. This helps us find topics that affect differing bias groups, if these bias groups react to the same real-world events, and what aspects of these events they react to. To achieve this we utilise a method where the Community Notes posted during a selected one-week period are taken and separated into different bias labels. For both of these groups, a set of a maximum of 30 keyphrases is generated using Keybert on a text merged from all of the Community Notes in a group. All of the keyphrases are then stemmed and if 2 of them have the same stem, they are merged and

[9] https://maartengr.github.io/KeyBERT/

their scores added. After this process, we keep the 10 highest-scoring keyphrases. In other parts of the work, this method is referred to as Temporal Topic Analysis.

5. Results

In this Section, we will review the results of the methods highlighted in the methodology chapter. For each section, we cover the results found by employing the corresponding method. In addition to the observations, each subsection includes visualisations of the findings.

5.1 Community Notes Visualization

Community Notes are a very popular feature, with the frequency of Community Notes being posted steadily climbing. We study the frequency of Community Notes in the time period between 28.01.2021 and 11.11.2023. Our observations as shown in Figure 3 indicate that during this period the frequency of left-leaning Community Notes being posted daily topped around 1000, while the frequency of right-leaning Community Notes being posted daily reached a high of around 200. This nearly 5x difference stays relatively the same for most weeks and also for aggregated data. It can be attributed to many factors with the first being the higher popularity of left-leaning news sources. The second possible reason is the young user base, with 3 in 4 Twitter users being under the age of 45 [53]. The first big spike on the graph, near October 2022 correlates with Elon Musk, who has been a vocal supporter of the feature acquiring Twitter (see Figure 4).

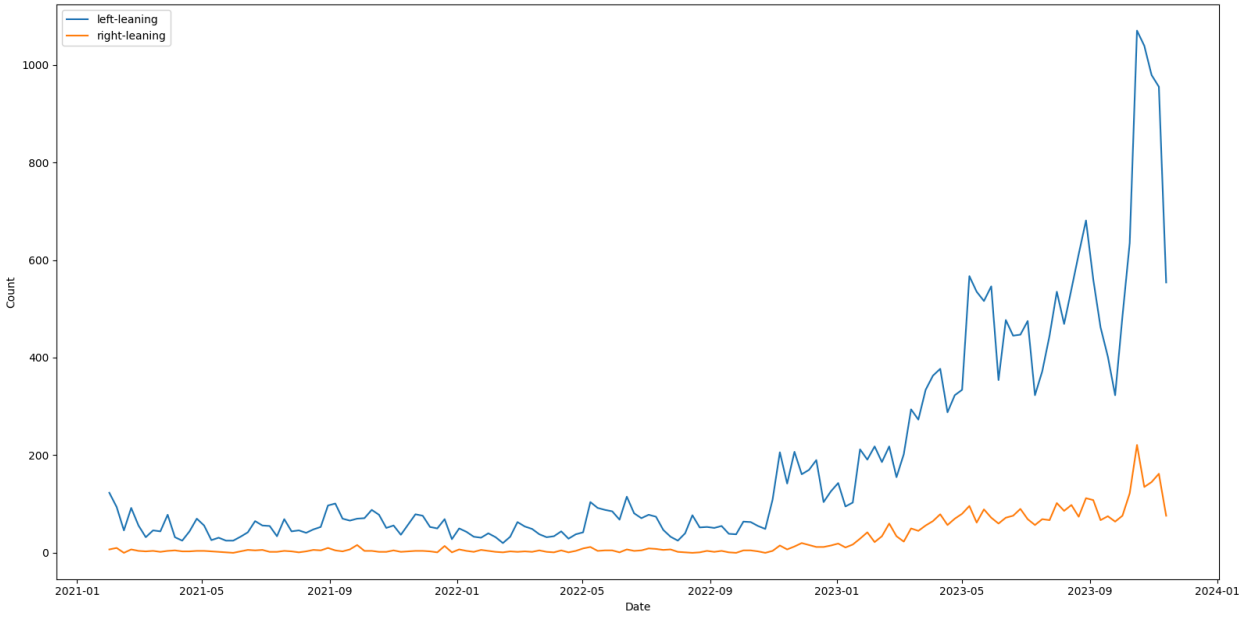

Figure 3. Change of Community Notes written by bias

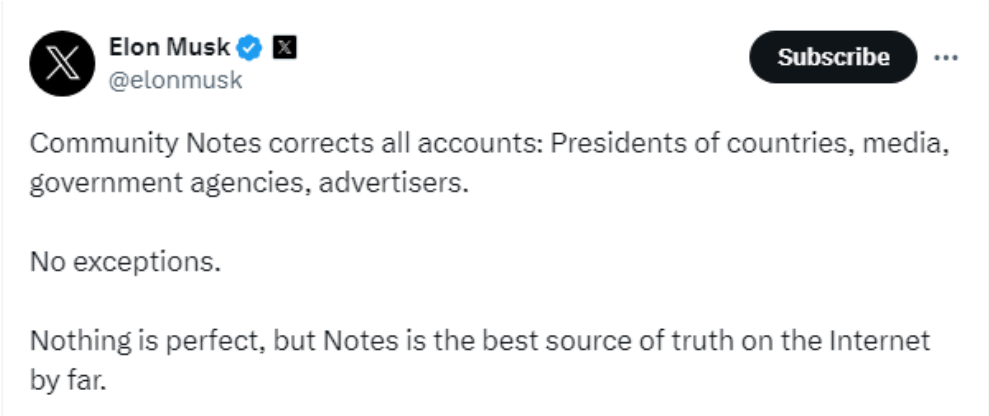

Figure 4: Public Endorsement of Community Notes by Elon Musk

5.2 Lexical analysis

POS Tagging: We POS Tag the Community Notes to find the average frequency of adjectives and the average frequency of words in Community Notes by bias group.

We show our observations for the analysis of the frequency of adjectives and other words in Figure 5, which represents a barplot of the frequency of words for each bias label such that the y-axis is the average frequency of words in a community note and the x-axis is the bias label. The total height of the bar represents the sum of adjectives and other words, so the average number of words in a Community Note. Our observations, as shown in Figure 5, indicate that there is no significant difference in the usage of words other than adjectives, with Right-Leaning Community Notes using the fewest frequency of words at 27.96 and Centre the greatest at 28.26, marking around a 1% difference. In the usage of adjectives, the differences are bigger, with Centre being the highest at 2.73, followed by Left-Leaning at 2.48 and Right-Leaning at 2.31. This marks around a 10% difference between Centre and Left-Leaning and around an 18% difference between Centre and Right-Leaning. This could indicate that Community Notes of Centre bias are more subjective, with previous works showing a correlation between the frequency of adjectives used and the subjectivity of the document [48].

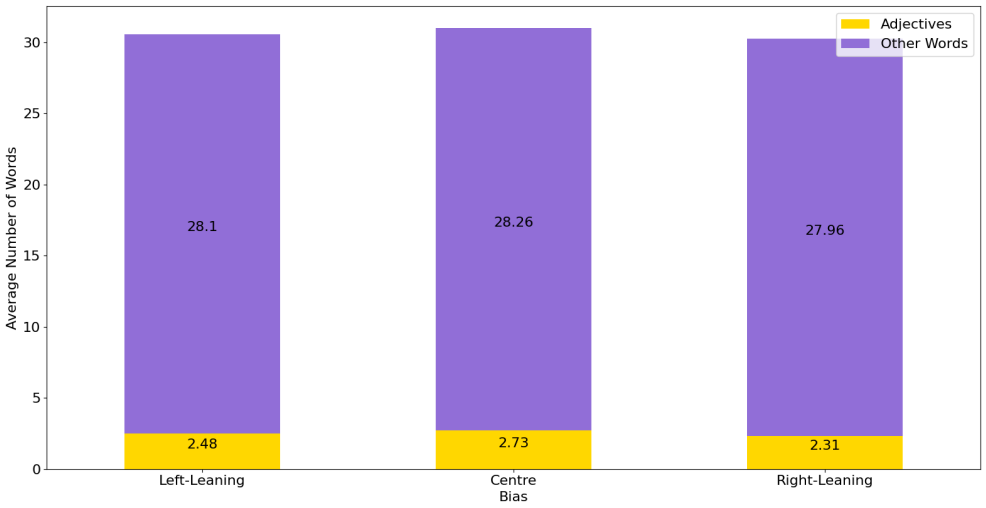

Figure 5. Average frequency of words and adjectives in Community Notes

Readability Analysis: We evaluate the readability of Community Notes using four different indexes, Dall-Chall, Flesch-Kincaid and Coleman-Liau Indices. We show our observations for the Dale-Chall index in Figure 6, which represents a box plot of the readability scores for each bias label such that the y-axis is the corresponding readability score and the x-axis is the bias label. Our observations, as shown in Figure 6, show no meaningful difference between the Dale-Chall scores of Community Notes with differing bias labels. The mean scores of left, centre and right-leaning Community Notes are 10.64, 10.83 and 10.51. The Q1 and Q3 scores also show marginal differences between Community Notes of separate bias groups.

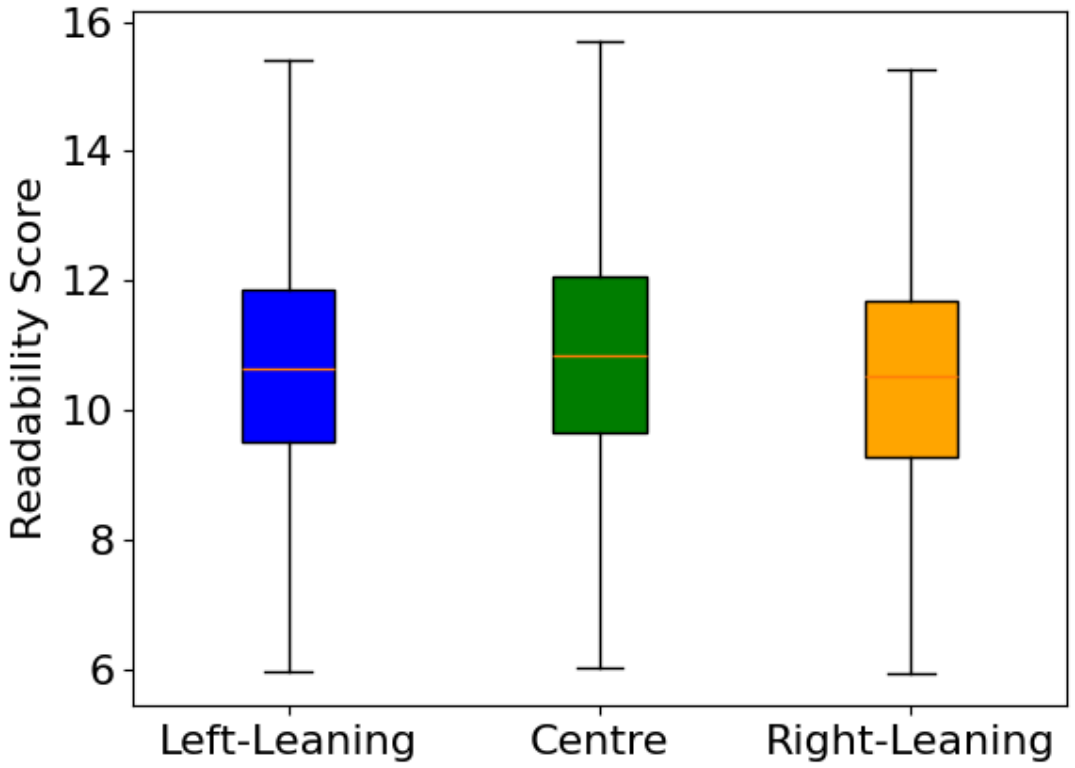

Figure 6. Box plot of the Dale-Chall readability score separated by bias

Our observations, as shown in Figure 7, suggest that right-leaning Community Notes have a considerably lower score, which indicates writing of greater difficulty to comprehend. The Q1 of right-leaning Community Notes is -85.67, while it is -62.61 for left-leaning and -57.68 for centre Community Notes. The difference is similar for the mean, with the corresponding scores being -30.96, -14.52, and -10.98 and also for Q3, with the scores being 3.66, 15.76, and 19.74.

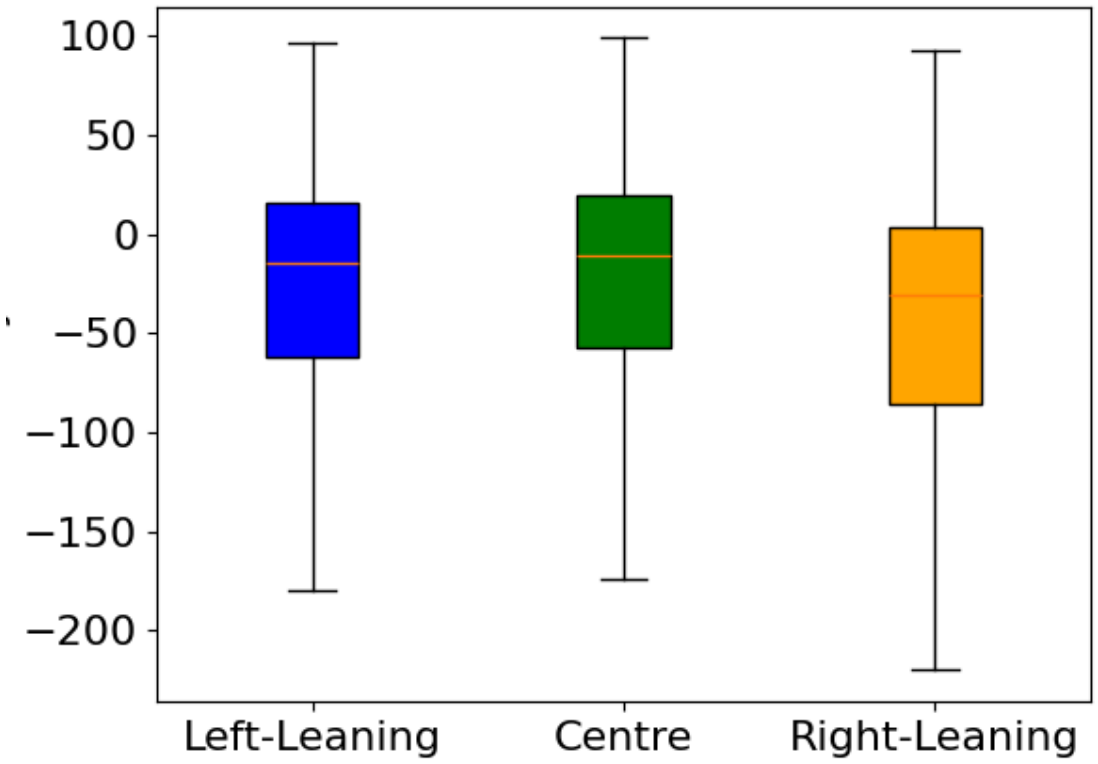

Figure 7. Box plot of the Flesch-Kincaid readability score separated by bias

The Coleman-Liau index, as shown in Figure 8, displays results similar to the Flesch-Kincaid with Community Notes of right-leaning bias having the highest scores which for this index indicates a higher level of difficulty. The biggest difference is in Q3, with the right-leaning Community Notes scoring 52.02, left-leaning 47.92 and centre 44.47.

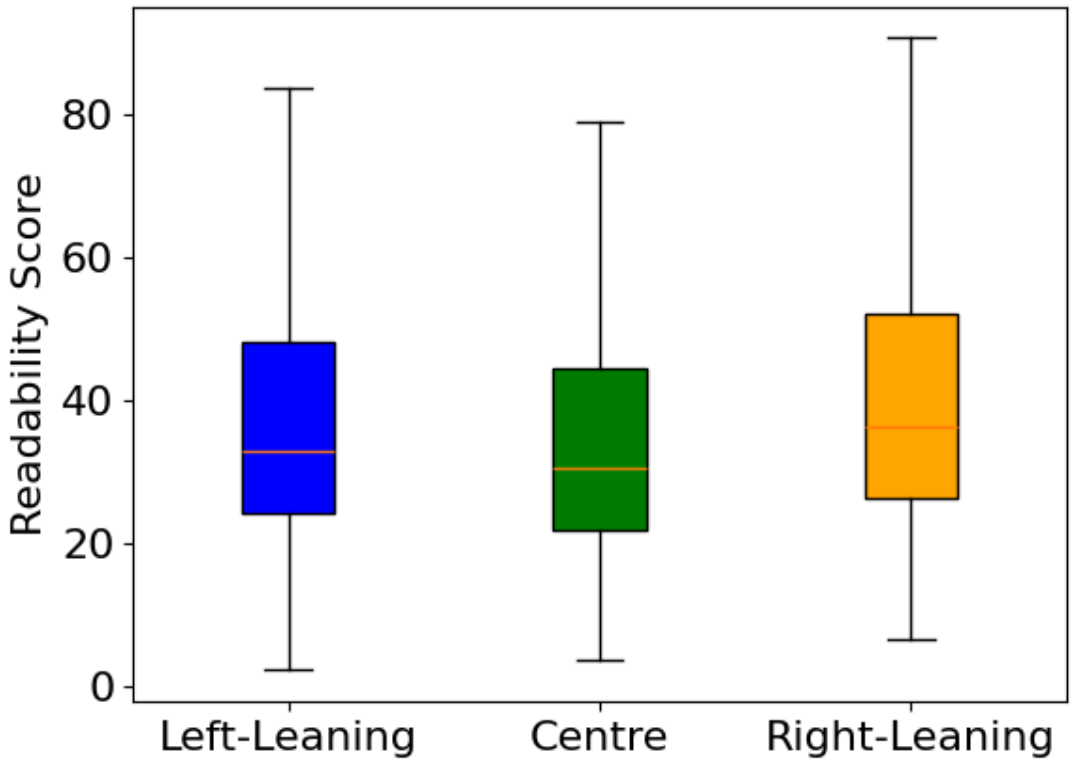

Figure 8. Box plot of the Coleman-Liau readability score separated by bias

5.3 Sentiment analysis

We utilise sentiment analysis in 3 different ways by finding word-level, sentence-level and note-level sentiment separately.

Word-level sentiment: We show our observations for the Word-level sentiment in Figure 9, which represents a barplot of the sentiment of words for each bias label such that the y-axis is the relative frequency of words with a given sentiment and the x-axis is the sentiment. Our observations as shown in Figure 9, indicate that the distribution of positive, negative, and neutral adjectives used is very similar for both left- and right-leaning Community Notes. There exists around a 1 percentage point gap in that left-leaning Community Notes use slightly more polarising words than the other groups, but the difference is not notable, and no conclusions can be drawn from it. Examples of adjectives with differing sentiment labels can be seen in Table 3.

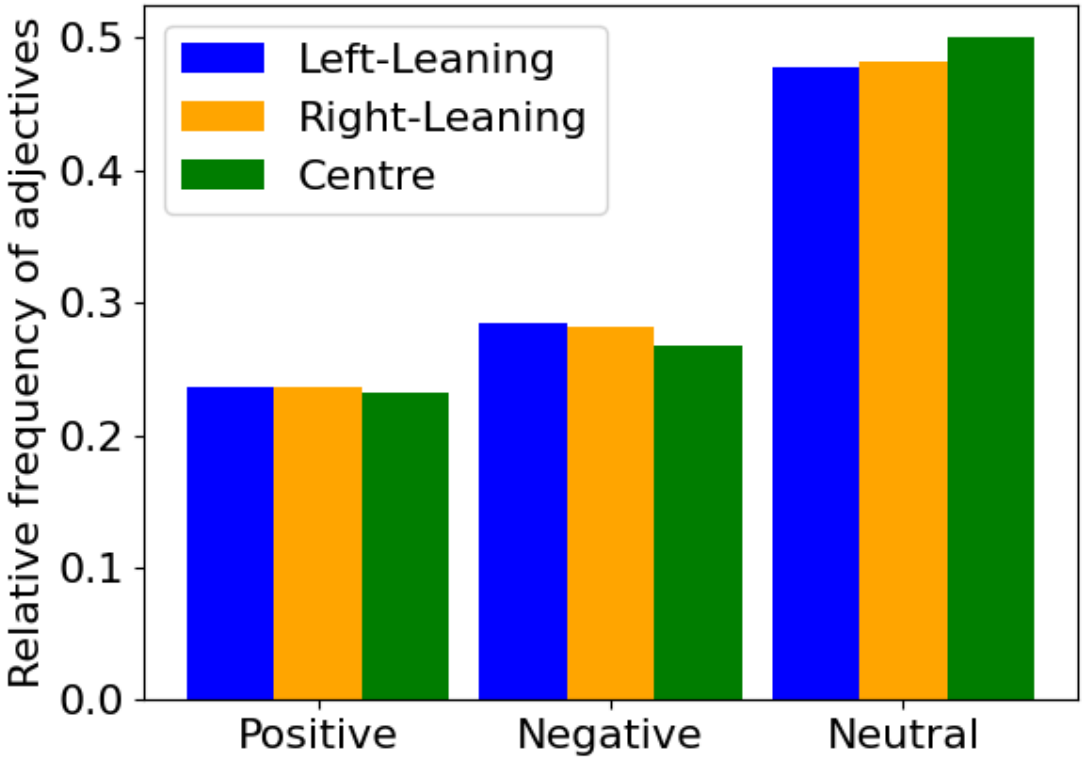

Figure 9. Distribution of adjectives by bias

Table 3. Examples of positive, negative, and neutral adjectives by bias

	Positive	Negative	Neutral
Left-leaning	functional	defamatory	vocal
Right-leaning	available	criminal	current
Centre bias	popular	fake	social

The results of the first sentence-sentiment approach, as explained in the methodology, can be observed in 10. The x-axis marks the bias of the corresponding column and the y-axis is the Community Note sentiment distribution based on the previously highlighted labelling, with the numbers on top of columns representing the sample size for the corresponding column. We also calculate and graph the z-score 99% confidence intervals in black at the top of the

bars, which we use to interpret the validity of our results. As shown in Figure 10, our results show that 40.3% of centre Community Notes, 44.7% of left-leaning Community Notes and 45% of right-leaning Community Notes are negative. This highlights a notable difference between the occurrence of negative sentiment in centre Community Notes when compared to left or right-leaning Community Notes.

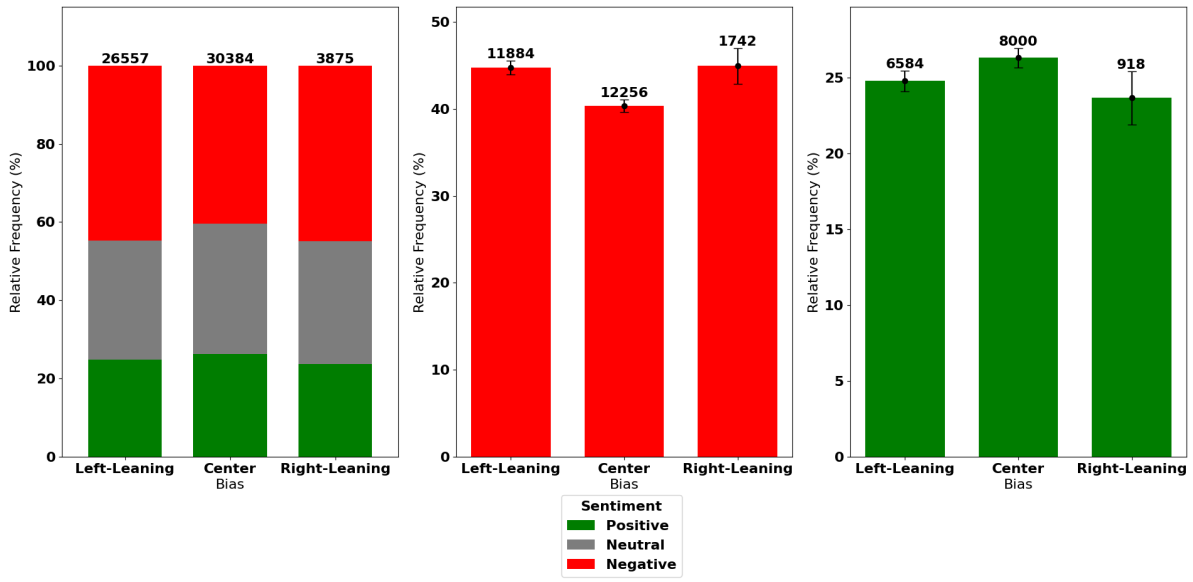

Figure 10. Sentiment distribution of bias groups based on sentence sentiment

For the second sentence-level sentiment approach, the individual sentence sentiment labels are used. The constructed graph has the same structure as Figure 11. The findings are similar to the previous graph, with the most notable difference again being the occurrence of negative sentiment. 24.8% of all sentences in centre Community Notes are negative, 27.5% in left-leaning Community Notes and 26.9% in right-leaning Community Notes.

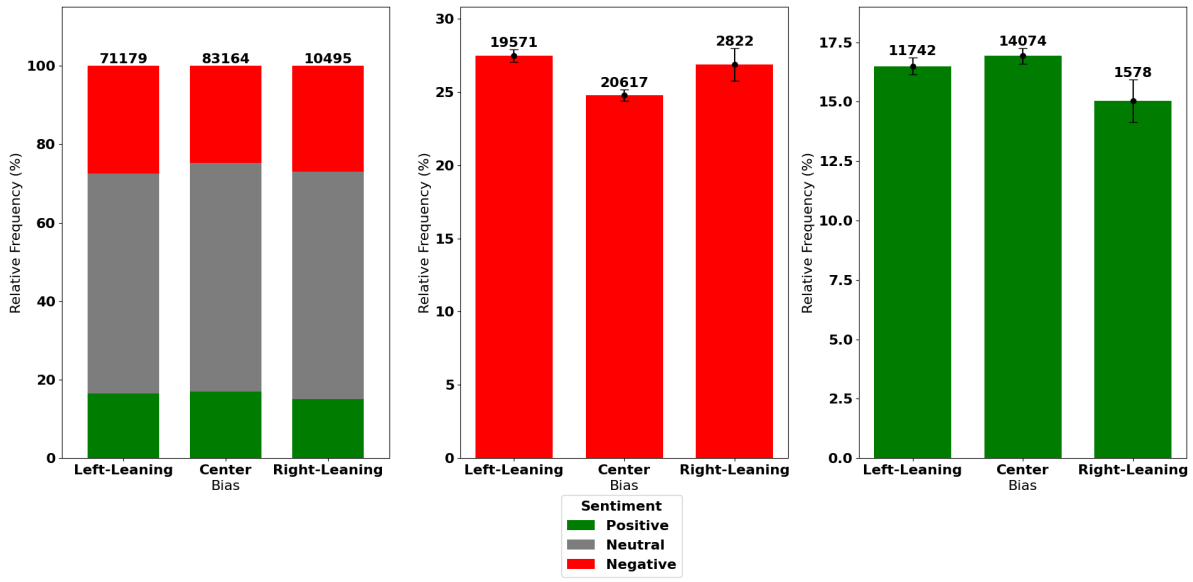

Figure 11. Distribution of sentence sentiment counts, grouped by bias

When comparing the note level sentiment between Community Notes of left, centre, and right bias it can be observed that left-leaning Community Notes tend to be more negative than centre Community Notes, with negative Community Notes making up 52.7% of all left-leaning Community Notes, while that number is only 46% for centre Community Notes. This is also meaningful as the sample sizes for both sets of Community Notes are large, as seen in Figure 6. Additionally, for all bias groups, only around 1% of Community Notes are positive, which can be explained by the nature of Community Notes, where adding context or debunking facts is usually done when the writer disagrees with the contents of a tweet and based on that has a negative attitude towards it.

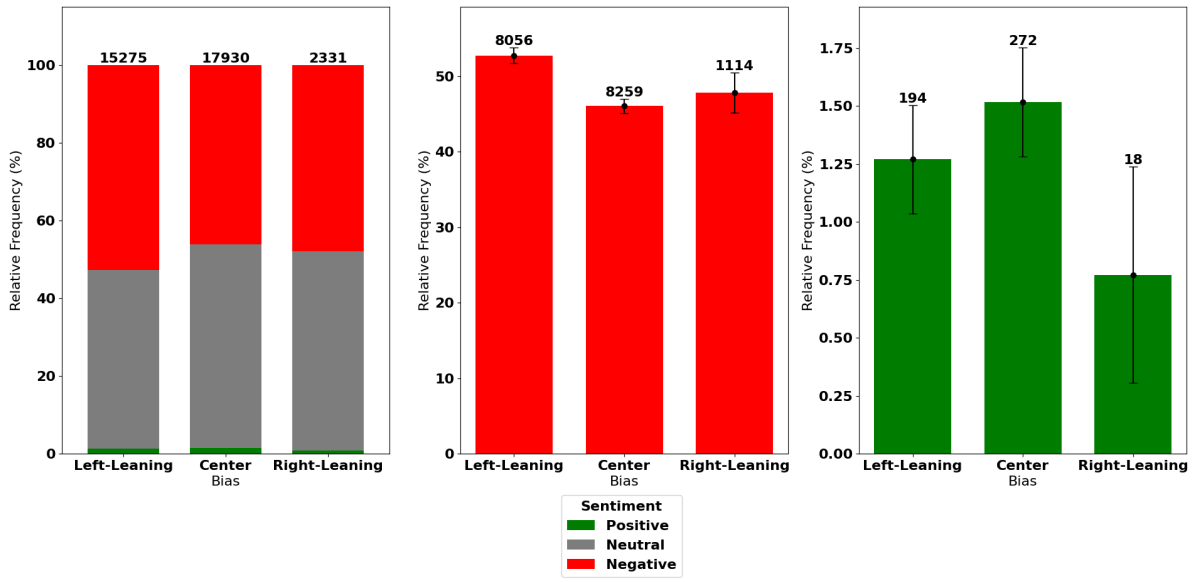

Figure 12. Distribution of Community Notes sentiment by bias

5.4 Sentiment temporal analysis

Only weeks with at least 10 Community Notes are utilised for these findings to reduce further the noise the low sample size introduces. The steady decline in polarity for left-leaning Community Notes can be attributed to an increase in the number of Community Notes, as seen in Figure 13. Right-leaning Community Notes, however, have a reverse trend of increasing polarity. This could be due to most weeks with at least 10 right-leaning Community Notes occurring after the Elon Musk takeover.

We also observe that during the period starting from 2023, when there is weekly data for both left and right-leaning Community Notes, the right-leaning Community Notes have a much higher standard deviation in terms of polarity. During this period, the standard deviation of polarity was 0.061 for right-leaning Community Notes and 0.0255 for left-leaning Community Notes. This can also be observed in Figure 13, as most of the data points for left-leaning Community Notes are very close to the linear regression line.

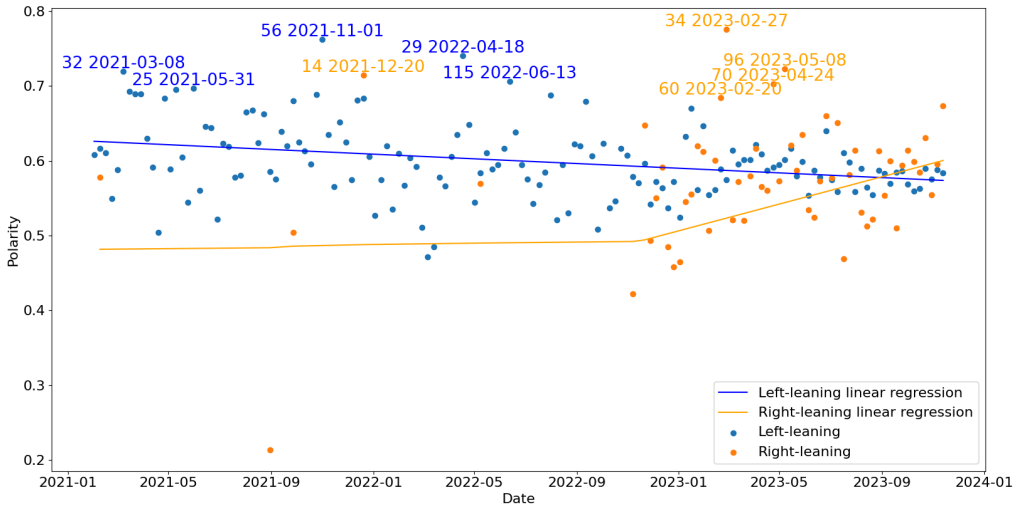

Figure 13. Change in Community Note polarisation over time

To find possible correlations between real-world events and spikes in polarity, we extract the keywords of a 1-week period, namely 06.06.2021 to 13.06.2022, using Period-Based Keyword Extraction. This period is chosen because it has the largest sample size among the top 5 highest polarity weeks for left-leaning Community Notes. The extracted keywords can be seen in Table 4.

Nine of the ten extracted keywords for left-leaning Community Notes are related to guns. This is most likely due to three events taking place during that week, which can all be attributed to the aftermath of the Uvalde school shooting, where 22 people were fatally shot [54]. The first of these events is Matthew McConaughey, a native of Uvalde, calling for gun reform at a white house briefing [55]. The second is a demonstration in support of gun control legislation called March for Our Lives, which took place in Washington DC and gathered thousands of attendees [56]. The third is U.S. Senator Chris Murphy announcing the Bipartisan Safer Communities Act, the first major gun legislation in nearly 30 years [57].

Among the top ten keywords for right-leaning Community Notes, only one is on the subject of guns. This could be due to the low sample size of seven Community Notes, of which one mentions firearms. It is also possible that this is due to right-leaning posters being more against gun control. Only 35% of Republicans, who are in the majority right-leaning, support stricter gun control laws, while 88% of Democrats, who are in the majority left-leaning, support these laws. The keyphrases containing “victims” and “violence” were manually

checked and are not related to the Uvalde school shooting. Any other keyphrases for right-leaning Community Notes do not correlate to any major real-world events.

Table 4. Top 10 keyphrases for left- and right-leaning Community Notes posted between 06.06.2021 and 13.06.2022

Left-Leaning keyphrase	**Keyphrase score**	**Right-Leaning keyphrase**	**Keyphrase score**
Weapon guns	0.8381	Russia reporter	0.5163
Carrying guns	0.7896	Pornography evidence	0.4511
Gun violence	0.5112	Russia appearing	0.4402
Confiscate weapons	0.4638	Victims stories	0.4293
Handgun bullet	0.4312	Child pornography	0.4055
News hearing	0.4312	Firearm knowing	0.3977
Guns day	0.4039	Reporter falsely	0.3854
Firearms brought	0.3846	Place russia	0.3810
Having gun	0.3843	Credibility victims	0.3768
Pistols bullet	0.3811	Violence survivors	0.3705

To further analyse relationships between real-world events and the polarity of Community Notes, another set of keyphrases is extracted using Period-Based Keyphrase Extraction. The chosen period is 01.05.2023 to 08.05.2023 due to it having the largest sample size among weeks with the top 5 mean polarity in right-leaning Community Notes. Keyphrases extracted from this time period can be observed in Table 4.

Two events compose eight out of ten keyphrases for the left-leaning Community Notes. One of these is an epidemiologist claiming that the US Teachers Union chief Randi Weingarten misrepresented a COVID study to the US Congress [58]. The other is the 2023 Allen, Texas mall shooting, where the perpetrator with the last name of Garcia fatally wounded 9 people [59]. This supports the previous findings that left-leaning fact-checks tend to target gun control more than right-leaning.

Six out of ten keyphrases for right-leaning Community Notes contain the word Trump, but no new major events concerning him took place during that period. The keyphrase "prince harry" occurs due to the news that Prince Harry took a commercial plane from the US to the UK [60]. For the other keyphrases, no corresponding events could be identified.

Table 5. Top 10 keyphrases for left- and right-leaning Community Notes posted between 01.05.2023 and 08.05.2023

Left-Leaning keyphrase	**Keyphrase score**	**Right-Leaning keyphrase**	**Keyphrase score**
Weingarten misrepresented	1.1219	Extensively trump	0.4098
Garcia evidence	0.8171	Racism incident	0.3856
Garcia investigated	0.4465	Trump fit	0.3852
Garcia suspected	0.4041	Piers morgan	0.3784
Unions weingarten	0.3923	Prince harry	0.3580
Garcia associated	0.3851	2016 trump	0.3441
Starting garcia	0.3764	Trump unworthy	0.3413
Controversy cnn	0.3743	Evidence racism	0.3353
Buzzfeed article	0.3740	Trump campaign	0.3338
Garcia perpetrator	0.3657	2018 trump	0.3311

Based on these two time periods, we observe that left-leaning Community Notes tend to react more to current news events. We also observe that left-leaning Community Notes are very vocal on the subject of gun violence, with 16 of 20 keyphrases for left-leaning being related to gun violence. The right-leaning Community Notes however are not as connected with current news.

Conclusions

In this thesis, we explored bias in Community Notes, a crowdsourced fact-checking solution. To our knowledge, only exploratory studies have been done on Community Notes, with none exploring their bias.

We collected all 323 382 Community Notes posted between 23.01.2021 and 11.11.2023. Three media monitoring sites were used to assign each Community Note a bias label. To assess the bias of Community Notes lexical features, sentiment analysis, temporal analysis and keyphrase extraction techniques were utilised.

The analysis of lexical features hints at Community Notes in the centre group using more adjectives when compared to other bias groups. Sentiment analysis results show a higher occurrence of negative Community Notes in the left-leaning group compared to the centre group. Analysing the keywords of 2 weeks with highly polarising word usage indicates that left-leaning Community Notes are more connected to real-world events, especially cases of gun violence, compared to right-leaning notes. These findings show promising results in identifying bias in Community Notes, with clear differences between bias groups being observed using the aforementioned methods.

In conclusion, we explored Community Notes to identify potential bias. Valuable insights have been gained regarding the sentiment and word usage in Community Notes by differing bias labels. This is the first work on this topic and further research is required for more conclusive findings. Possible future approaches could leverage LLMs for bias detection in Community Notes, analyse the demographics of Community Note posters and how bias affects the acceptance of Community Notes.

References

[1] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR). 2021 Jul 13. doi: 10.1145/3457607.

[2] Oxford Advanced Learner's Dictionary [Internet]. Oxford: Oxford University Press; c2024. bias noun. [cited 2024 May 9]. Available from: https://www.oxfordlearnersdictionaries.com/definition/english/bias_1

[3] Baron DP. Persistent media bias. Journal of Public Economics. 2006 Jan 1;90(1-2). doi:10.1016/j.jpubeco.2004.10.006

[4] Chakraborty A, Messias J, Benevenuto F, Ghosh S, Ganguly N, Gummadi K. Who makes trends? understanding demographic biases in crowdsourced recommendations. In Proceedings of the International AAAI Conference on Web and Social Media. 2017 May 3;11(1). doi:10.1609/icwsm.v11i1.14894

[5] Hamborg F, Donnay K, Gipp B. Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal on Digital Libraries. 2019 Dec;20(4). doi:10.1007/s00799-018-0261-y

[6] Ribeiro F, Henrique L, Benevenuto F, Chakraborty A, Kulshrestha J, Babaei M, Gummadi K. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In Proceedings of the International AAAI Conference on Web and Social Media. 2018 Jun 15;12(1). doi:10.1609/icwsm.v12i1.15025.

[7] Spinde T, Kreuter C, Gaissmaier W, Hamborg F, Gipp B, Giese H. Do you think it's biased? how to ask for the perception of media bias. In 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). 2021 Sep 27. IEEE. doi: 10.1109/JCDL52503.2021.00018

[8] Mitchell A, Gottfried J, Barthel M, Shearer E. The Modern News Consumer [Internet]. Pew Research Center's Journalism Project. 2016. [cited 2024 May 9]. Available from: https://www.pewresearch.org/journalism/2016/07/07/the-modern-news-consumer/

[9] Allcott H, Gentzkow M. Social media and fake news in the 2016 election. Journal of economic perspectives. 2017 May 1;31(2). doi: 10.1257/jep.31.2.211.

[10] Duseja N, Jhamtani H. A sociolinguistic study of online echo chambers on twitter. In Proceedings of the third workshop on natural language processing and computational social science. 2019 Jun. doi: 10.18653/v1/W19-2109.

[11] Dixon L, Li J, Sorensen J, Thain N, Vasserman L. Measuring and mitigating unintended

bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018 Dec 27. doi:10.1145/3278721.3278729

[12] Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. The echo chamber effect on social media. Proceedings of the National Academy of Sciences. 2021 Mar 2;118(9). doi: 10.1073/pnas.2023301118.

[13] Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG. Shifting attention to accuracy can reduce misinformation online. Nature. 2021 Apr 22;592(7855). doi:10.1038/s41586-021-03344-2

[14] Mozafari M, Farahbakhsh R, Crespi N. Hate speech detection and racial bias mitigation in social media based on BERT model. PloS one. 2020 Aug 27;15(8). doi: 10.1371/journal.pone.0237861

[15] MacAvaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O. Hate speech detection: Challenges and solutions. PloS one. 2019 Aug 20;14(8). doi: 10.1371/journal.pone.0221152

[16] Amazeen MA. Journalistic interventions: The structural factors affecting the global emergence of fact-checking. Journalism. 2020 Jan;21(1). doi:10.1177/1464884917730217

[17] Vargo CJ, Guo L, Amazeen MA. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. New media & society. 2018 May;20(5). doi:10.1177/1461444817712086

[18] Chuai Y, Tian H, Pröllochs N, Lenzini G. The Roll-Out of Community Notes Did Not Reduce Engagement With Misinformation on Twitter. arXiv preprint arXiv:2307.07960. 2023 Jul 16. doi: 10.48550/arXiv.2307.07960.

[19] How to read and share Notes | X Help [Internet]. help.twitter.com. [cited 2024 May 9]. Available from: https://help.twitter.com/en/using-x/notes

[20] Community Notes. Notes that cite sources have a much higher likelihood of earning a ‘Helpful’ status. This is not surprising, as sources make notes more easily verifiable by viewers and raters. Starting today, sources are now required for proposed notes. We haven’t previously required this: https://twitter.com/CommunityNotes/status/1714409083036643549. 2023 Oct 18 [cited 2024 May 9] [Tweet]. Available from: https://twitter.com/CommunityNotes

[21] X. About Community Notes on X | X Help [Internet]. help.twitter.com. [cited 2024 May 9]. Available from: https://help.twitter.com/en/using-x/community-notes

[22] Notes shown on X [Internet]. communitynotes.x.com. [cited 2024 May 9]. Available

from: https://communitynotes.x.com/guide/en/contributing/notes-on-twitter

[23] Evaluation [Internet]. communitynotes.x.com. [cited 2024 May 9]. Available from: https://communitynotes.x.com/guide/en/under-the-hood/guardrails

[24] Renault T, Amariles DR, Troussel A. Collaboratively adding context to social media posts reduces the sharing of false news. arXiv preprint arXiv:2404.02803. 2024 Apr 3. doi:10.48550/arXiv.2404.02803

[25] Devinney H, Björklund J, Björklund H. Theories of "gender" in nlp bias research. In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. 2022 Jun 21. doi:10.1145/3531146.3534627

[26] Groseclose T, Milyo J. A measure of media bias. The quarterly journal of economics. 2005 Nov 1;120(4). doi:10.1162/003355305775097542

[27] Spinde T, Rudnitckaia L, Mitrović J, Hamborg F, Granitzer M, Gipp B, Donnay K. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. Information Processing & Management. 2021 May 1;58(3). doi:10.1016/j.ipm.2021.102505

[28] Sen A, Ghatak D, Khanuja G, Rekha K, Gupta M, Dhakate S, Sharma K, Seth A. Analysis of media bias in policy discourse in india. In Proceedings of the 5th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies. 2022 Jun 29. doi:10.1145/3530190.3534798

[29] Nye JS, Alterman E. What liberal media?: The truth about bias and the news. Basic Books; 1990.

[30] Babaei M, Kulshrestha J, Chakraborty A, Redmiles EM, Cha M, Gummadi KP. Analyzing biases in perception of truth in news stories and their implications for fact checking. IEEE Transactions on Computational Social Systems. 2021 Jul 28;9(3). doi:10.1109/TCSS.2021.3096038.

[31] An J, Cha M, Gummadi K, Crowcroft J, Quercia D. Visualizing media bias through Twitter. In Proceedings of the International AAAI Conference on Web and Social Media. 2012;6(2). doi:10.1609/icwsm.v6i2.14343.

[32] Chakraborty R, Bhavsar M, Dandapat SK, Chandra J. Tweet summarization of news articles: An objective ordering-based perspective. IEEE Transactions on Computational Social Systems. 2019 Jul 24;6(4). doi:10.1109/TCSS.2019.2926144.

[33] DiGrazia J, McKelvey K, Bollen J, Rojas F. More tweets, more votes: Social media as a quantitative indicator of political behavior. PloS one. 2013 Nov 27;8(11). doi:10.1371/journal.pone.0079449.

[34] An J, Cha M, Gummadi K, Crowcroft J. Media landscape in Twitter: A world of new conventions and political diversity. In Proceedings of the International AAAI Conference on Web and Social Media. 2011;5(1). doi:10.1609/icwsm.v5i1.14118.

[35] Cha M, Haddadi H, Benevenuto F, Gummadi K. Measuring user influence in twitter: The million follower fallacy. InProceedings of the international AAAI conference on web and social media. 2010 May 16;4(1). doi:10.1609/icwsm.v4i1.14033.

[36] Vosoughi S, Roy D, Aral S. The spread of true and false news online. science. 2018 Mar 9;359(6380). doi:10.1126/science.aap9559

[37] Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W. The spreading of misinformation online. Proceedings of the national academy of Sciences. 2016 Jan 19;113(3). doi:10.1073/pnas.1517441113.

[38] Chen W, Pacheco D, Yang KC, Menczer F. Neutral bots probe political bias on social media. Nature communications. 2021 Sep 22;12(1). doi:10.1038/s41467-021-25738-6

[39] Pierri F, Luceri L, Jindal N, Ferrara E. Propaganda and misinformation on Facebook and Twitter during the Russian invasion of Ukraine. In Proceedings of the 15th ACM web science conference 2023. 2023 Apr 30. doi:10.1145/3578503.3583597.

[40] Schuster T, Shah D, Yeo YJ, Ortiz DR, Santus E, Barzilay R. Towards Debiasing Fact Verification Models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Nov. doi:10.18653/v1/D19-1341

[41] Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. InProceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Jun;1. doi:10.18653/v1/N18-1074

[42] Augenstein I, Lioma C, Wang D, Lima LC, Hansen C, Hansen C, Simonsen JG. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. InProceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Nov. doi:10.18653/v1/D19-1475

[43] Pilarski M, Solovev K, Pröllochs N. Community Notes vs. Snoping: How the Crowd Selects Fact-Checking Targets on Social Media. arXiv preprint arXiv:2305.09519. 2023 May 16. doi:10.48550/arXiv.2305.09519.

[44] How to post – what is a post, keyboard shortcuts, and sources | X Help [Internet]. help.twitter.com. [cited 2024 May 9]. Available from:

https://help.twitter.com/en/using-x/how-to-post

[45] Downloading data [Internet]. communitynotes.x.com. [cited 2024 May 9]. Available from: https://communitynotes.x.com/guide/et/under-the-hood/download-data

[46] Harper's BAZAAR. King Charles Takes Military Role from Prince Harry and Gives it to William: https://twitter.com/harpersbazaarus/status/1787909949600809293. 2024 May 7 [cited 2024 May 9] [Tweet]. Available from: https://twitter.com/harpersbazaarus

[47] Sarica S, Luo J. Stopwords in technical language processing. Plos one. 2021 Aug 5;16(8). doi:10.1371/journal.pone.0254937

[48] Rittman R, Wacholder N, Kantor P, Ng KB, Strzalkowski T, Sun Y. Adjectives as indicators of subjectivity in documents. Proceedings of the American Society for Information Science and Technology. 2004;41(1). doi:10.1002/meet.1450410141.

[49] Klare GR. The measurement of readability: useful information for communicators. ACM Journal of Computer Documentation (JCD). 2000 Aug 1;24(3). doi:10.1145/344599.344630

[50] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal. 2014 Dec 1;5(4). doi:10.1016/j.asej.2014.04.011.

[51] Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. InProceedings of the international AAAI conference on web and social media. 2014 May 16;8(1). doi:10.1609/icwsm.v8i1.14550.

[52] Tong Z, Zhang H. A text mining research based on LDA topic modelling. InInternational conference on computer science, engineering and information technology. 2016 May 6. doi:10.5121/csit.2016.60616.

[53] Koray E. Ipsos Connect Tech Tracker Q4 2015. Ipsos. 2015 Dec 15. Available from: https://www.ipsos.com/sites/default/files/publication/1970-01/Ipsos_Connect_Tech_Tracker_Q4_2015.pdf

[54] Gamboa S, Romero D, Burke M, Fieldstadt E. 19 children, 2 teachers killed in shooting at Robb Elementary School in Uvalde, Texas. NBC News [Internet]. 2022 May 24 [cited 2024 May 9]. Available from: https://www.nbcnews.com/news/us-news/police-respond-active-shooter-texas-elementary-school-rcna30339

[55] Franck T. Actor Matthew McConaughey calls for gun reform in White House briefing. CNBC [Internet]. 2022 Jun 7 [cited 2024 May 9]. Available from: https://www.cnbc.com/2022/06/07/watch-live-actor-matthew-mcconaughey-joins-white-house-press-briefing-on-guns.html

[56] Silverman E, Wan W, Hilton J, Lumpkin L, Cox E, Ruane ME, Latson S. March for Our Lives 2022: Thousands gather to protest gun violence. Washington Post [Internet]. 2022 Jul 1 [cited 2024 May 9]. Available from: https://www.washingtonpost.com/dc-md-va/2022/06/11/march-for-our-lives-dc-protests/

[57] Chris Murphy. 🚨NEWS: We have a deal. Today a bipartisan group of 20 Senators (10 D and 10 R) is announcing a breakthrough agreement on gun violence - the first in 30 years - that will save lives. I think you'll be surprised at the scope of our framework: https://twitter.com/ChrisMurphyCT/status/1536013602846560256. 2022 Jun 12 [cited 2024 May 9] [Tweet]. Available from: https://twitter.com/ChrisMurphyCT

[58] Christenson J. Randi Weingarten misrepresented COVID study to Congress, author claims. New York Post [Internet]. 2023 Apr 28 [cited 2024 May 9]. Available from: https://nypost.com/2023/04/28/covid-study-doc-randi-weingarten-misrepresented-my-work/

[59] Romero D. Dallas-area mall shooting: 9 dead including suspect, 3 in critical condition. NBC News [Internet]. 2023 May 7 [cited 2024 May 9]. Available from: https://www.nbcnews.com/news/us-news/shooting-reported-dallas-area-outlet-mall-rcna83220

[60] Ng K. Prince Harry stuns passengers as he arrives on commercial flight for King Charles III's coronation. The Independent [Internet]. 2023 May 6 [cited 2024 May 9]. Available from: https://www.independent.co.uk/life-style/royal-family/prince-harry-coronation-travel-commercial-flight-b2333691.html

Appendix

I. Source code

The analysis for this work was performed using Python. All of the used code is available on GitHub and can be accessed here: https://github.com/monitor640/CommunityNotesBias

II. License

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Simon Fox Kuuse,

(author's name)

1. grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, my thesis

Studying bias in Twitter (X) Community Notes,

(title of thesis)

supervised by Uku Kangur, Roshni Chakraborty.

(supervisor's name)

2. I grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in points 1 and 2.

4. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Simon Fox Kuuse

15/05/2024