CLAIM SENSING: A STUDY LINKING FACTUAL CLAIMS TO HUMAN BEHAVIORS ON SOCIAL MEDIA

by

ZEYU ZHANG

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2024

ABSTRACT

CLAIM SENSING: A STUDY LINKING FACTUAL CLAIMS TO HUMAN BEHAVIORS ON SOCIAL MEDIA

ZEYU ZHANG, Ph.D.

The University of Texas at Arlington, 2024

Supervising Professor: Dr. Chengkai Li

The ubiquity of social media has transformed it into a rich source for reflecting people's opinions, behaviors, and interactions. Users frequently encounter factual claims in news, stories, and political statements, which can be either true or false. These claims significantly shape people's minds and behaviors, influencing not only individual perspectives but also broader public discourse. This study explores individuals' behaviors and perceptions toward factual claims by leveraging the concept of "check-worthiness" to analyze the relationship between such claims and user behaviors across datasets containing tens of millions of social media posts, particularly tweets from the platform X (formerly Twitter). It addresses key questions, including whether users exhibit different posting tendencies based on the check-worthiness of claims, the underlying reasons for these differences, and whether users are more likely to engage with content that aligns with the check-worthiness levels of their own posts.

Furthermore, the research introduces Wildfire, an innovative social sensing platform that empowers laypersons to conduct social sensing tasks using Twitter data without requiring programming or data analytics skills. Unlike existing tools that

rely on simple keyword-based searches, Wildfire employs a heuristic graph exploration method to selectively expand the collected tweet-account graph, enhancing the collection of task-relevant data. This platform also offers a range of analytic tools, such as text classification, topic generation, and entity recognition, facilitating tasks such as trend analysis and public opinion sensing.

In addition to these methodological advancements, the research includes several real-world case studies that contribute to understanding the surveillance and impact of factual claims on specific topics, particularly in the contexts of the COVID-19 pandemic and climate change discussions on social media. Utilizing large language models, the study matches tweets with curated facts and misinformation, analyzing their stances and spatio-temporal spread. The findings highlight trends in misinformation during the COVID-19 pandemic and reveal the public's general tendency to believe claims regardless of their veracity.

In summary, the integration of social media into our daily lives has profoundly transformed how we interact with information, necessitating sophisticated approaches to understanding and managing the vast array of factual claims circulating online. The rise of platforms like Twitter has amplified the impact of these factual claims on public opinion and major societal events. By adopting the innovative approach of claim sensing, this study aims to bridge the gap between factual claims and human behavior.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Social media has become an integral part of our lives, reflecting people's opinions, behaviors, and interactions. It provides a valuable means to observe and interpret societal phenomena and discover insights about our society. In our present era, an unprecedented surge of dissemination of assertions has taken root within our society. The digital worlds, notably prominent platforms such as Twitter (now called X), have become a breeding ground for various factual claims in news, stories, and persuasive statements. These factual claims significantly influence public opinion, affecting the outcomes of major events such as legislation [1, 2] and presidential elections [3, 4, 5, 6]. Consequently, we have witnessed a significant proliferation of fact-checking endeavors globally. Numerous researchers and experts are engaged in diverse works concerning factual claims, which encompass statements based on verifiable information. These efforts span activities such as detecting [7], tracking [8], and evaluating factual claims [9]. The practice of fact-checking has evolved into a pivotal interdisciplinary field, commanding attention across a spectrum of areas such as computer science, journalism, and communication.

While many researchers have delved into the realm of fact-checking and factual claims, a notable knowledge gap persists in our understanding of how factual claims wield influence over people's interactions on social media. Observations and explanations of people's behaviors pertaining to factual claims are needed in order to fill this gap. This would entail answering many crucial questions, including whether individuals exhibit different behavioral tendencies toward factual claims and the underlying

factors that drive and differentiate these tendencies. Moreover, can we apply the age-old adage "Birds of a feather flock together" to denizens of social media, particularly concerning their responsiveness to factual claims? These unknowns offer us a new perspective to study social media. The answers to these unknowns may facilitate studies such as behavioral modeling and recommendation systems by providing noteworthy new features. Moreover, it may foster research in fields such as psychology and sociology by introducing new human behavioral patterns on social media. Therefore, we leveraged the concept of "check-worthiness" to observe people's behavioral tendencies toward factual claims and gathered data from Twitter to conduct various statistical analyses, which led to meaningful results presented in Chapter 3.

The above practice offered us a new perspective on understanding and interpreting human society. We refer to this approach as "Claim Sensing"——observing and interpreting phenomena to uncover insights about human behaviors by collecting, processing, and analyzing factual claims. More broadly, the practice of gathering and analyzing perspectives from social media and internet communications is known as social sensing [10]. These practices ideally need a large volume of data and various analytical tools to sense one or more perspectives of certain populations. Nowadays, collecting desired data can be expensive and complicated due to more and more restricted data policies and technical barriers, especially for laypersons. Therefore, we designed and implemented a social sensing platform, called Wildfire, to support users in conducting social sensing tasks using Twitter data without programming and data analytics skills. Existing open-source and commercial social sensing tools only support data collection using simple keyword-based or account-based search. On the contrary, Wildfire employs a heuristic graph exploration method to selectively expand the collected tweet-account graph in order to further retrieve more task-relevant tweets and accounts. This approach allows for the collection of data to support complex social

sensing tasks that cannot be met with a simple keyword search. In addition, Wildfire provides a range of analytic tools, such as text classification, topic generation, and entity recognition, which can be crucial for tasks such as trend analysis. The platform also provides a web-based user interface for creating and monitoring tasks, exploring collected data, and performing analytics. The details of Wildfire will be presented in Chapter 4.

By leveraging the knowledge and tools of claim sensing, we can address various sociological questions and challenges by observing and analyzing people's behaviors concerning factual claims. For example, one might seek to monitor misinformation and debunk its spread on social media or understand public opinion for selected topics and events. This study presents several real-world cases in Chapter 5, including understanding the surveillance and impact of factual claims on specific topics, particularly in the contexts of the COVID-19 pandemic and climate change discussions on social media.

This study opens the door to a multitude of intriguing questions that could significantly impact various fields. How do factual claims influence human behavior, and what patterns emerge in public opinion when factual claims are present? Can we predict behavioral trends based on individuals' interactions with factual claims, and how might this reshape our approach to social media analytics? By providing new tools and methodologies, this research not only deepens our understanding of digital interactions but also offers practical applications in combating misinformation, enhancing public discourse, and fostering a more informed society. The answers to these questions could have far-reaching implications, potentially transforming how we perceive and engage with the digital world. This research could also advance fields such as computational social science, communication studies, and public policy, ultimately helping to shape more accurate models of human behavior in digital spaces.

CHAPTER 2

BACKGROUND

There is a significant body of work related to claim sensing, with fact-checking being one of the most pertinent tasks. Fact-checking is a critical need in journalism, supported by dedicated institutions such as Politifact, [1] Snopes, [2] and FactCheck.org. [3] These organizations employ numerous fact-checkers to verify various claims, a process that is both costly and time-consuming. For example, evaluating a claim such as the one in Figure 2.1 requires searching through potentially many sources, assessing the reliability of each, and then deriving an accurate conclusion. This can take professional fact-checkers several hours or even days [11, 12]. The challenge is further compounded by tight deadlines, particularly within internal processes [13], and studies suggest that fewer than half of all published articles undergo verification [14]. Given the vast influx of new information and the speed at which it spreads, manual validation is insufficient. As a result, researchers are exploring methods to achieve automated fact-checking.

Automated fact-checking has been a topic of discussion in the computational journalism community [15, 16]. This area has also garnered attention from other fields, including artificial intelligence and natural language processing (NLP). Automated fact-checking is a complex task that encompasses various challenging subtasks, such as identifying factual claims, searching for relevant knowledge and data, gathering evidence, and generating both the verdict and an explanation. Each of these subtasks

[1] http://www.politifact.com

[2] http://www.snopes.com

[3] http://www.factcheck.org

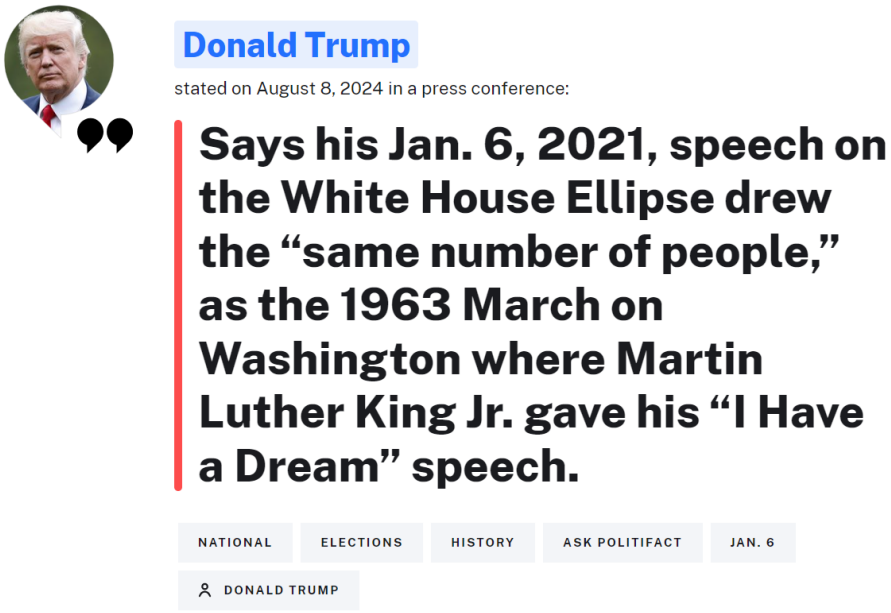


Figure 2.1: Example of a claim to fact-check

is demanding in its own right, which is why researchers often focus on individual components rather than attempting to achieve automated fact-checking as a singular, unified task. Several studies have surveyed research focusing on different aspects of automated fact-checking. Zubiaga et al. [17] and Islam et al. [18] focus on identifying rumors on social media. Küçük and Can [19] and Hardalov et al. [20] spent their efforts on detecting the stance of a given piece of evidence towards a claim. Kotonya and Toni [21] examine the generation of explanations and justifications for fact-checks. Nakov et al. [22] provide a survey of automated methods designed to assist human fact-checkers.

There are also some tasks not directly for fact-checking but highly related to it. For example, fake news detection and analysis. Fake news detection has a different scope than fact-checking, as it focuses on evaluating news articles and may involve labeling items based on factors unrelated to their truthfulness, such as identifying

sarcasm [23, 24]. Additionally, considerations such as the audience targeted by the claim, the intentions behind it, and its presentation style are often taken into account. These aspects are also relevant in the context of propaganda detection, which was recently reviewed by Da San Martino et al. [25]. In contrast, the studies of fact-checking primarily focus on verifying the truthfulness of general-domain claims. Lazer et al. [26] and Zhou et al. [24] reviewed research on fake news, encompassing both descriptive studies on the issue and efforts to combat it using computational methods. Additionally, Oshikawa et al. [23] provided an extensive overview of NLP approaches to fake news detection. Shu et al. [27] and da Silva et al. [28] focused on research related to fake news detection and fact-checking, particularly within the context of social media data.

In addition to research on automated fact-checking in technology, there are studies examining public opinion and responses to fact-checking. Rich et al. [29] explored whether the public supports fact-checking social media content, particularly from politicians. Brandtzaeg et al. [30] investigated how social media users view online fact-checking and verification services. Zhang et al. [31] analyzed the impact of fact-checking vaccine misinformation on social media on people's attitudes toward vaccines. Jiang et al. [32] examined the linguistic signals, especially emotional and topical ones, found in user comments when misinformation and fact-checking are present. Clayton et al. [33] assessed the effectiveness of fact-checking strategies employed by Facebook and other social media platforms to combat false stories. A number of similar studies delve into people's perceptions and reactions to fact-checking [34, 35, 36, 37].

In addition to research focused on general aspects of fact-checking, many studies focus on fact-checks or factual claims related to specific topics or events. For instance, Carey et al. [38] examined the short-term impact of fact-checks on COVID-19 miscon-

ceptions in the U.S., Great Britain, and Canada. Wintersieck et al. [39] explored how fact-checking affects people's attitudes and evaluations of political candidates. Newell et al. [40] studied whether consumers who view an advertisement with a misleading environmental claim develop significantly different attitudes toward the ad compared to those who see a similar, truthful ad. Van der Meer et al. [41] investigated misinformation treatment in public health by analyzing the effects of different types and sources of corrective information. There are plenty of similar studies, which can be associated with the concept of claim sensing, although they have not been categorized as such before. In this research, we propose the concept of claim sensing and explicitly link factual claims with human behaviors. By doing so, we aim to offer a new perspective for observing human behaviors and to inspire further insights in related social science studies.

CHAPTER 3

EXPLORING BEHAVIORAL TENDENCIES ON SOCIAL MEDIA: A PERSPECTIVE THROUGH CLAIM CHECK-WORTHINESS

3.1 Introduction

As we mentioned in Chapter 1, nowadays, digital platforms like Twitter and Facebook have become production centers for factual claims. Numerous researchers and experts are currently engaged in diverse works concerning factual claims. Despite extensive research on fact-checking and factual claims, a significant gap remains in understanding how these claims with varying significance influence social media interactions. To address this, it's essential to explore whether individuals possess behavioral tendencies toward factual claims and the factors driving these behaviors.

To study this subject, a suitable instrument is necessary to measure the strength or importance of a claim. Existing research provides a valuable resource. Within the field of automated fact-checking, extracting claims that are objectively fact-checkable makes the task more amenable to automation while reducing the volume of content needing manual fact-checking. More specifically, researchers have forged invaluable tools to detect check-worthy factual claims [42, 43, 44, 45, 46]. These efforts furnish a consistent yardstick for evaluating the importance of a claim to be fact-checked—denoted as "check-worthiness." As stated in [11], the initial work defined the concept of check-worthiness, check-worthy claims are those of which the general public would be interested in knowing the veracity—whether the claims are true or false. For instance, as displayed in Figure 3.1, (a) depicts a claim of significant check-worthiness, as it is highly probable that the public is interested in its veracity. In contrast,

(a) Claim with high check-wothiness

(b) Claim with relatively low check-wothiness

(c) Claim with low check-wothiness

Figure 3.1: Examples of claims with different check-worthiness levels

(b) conveys relatively low check-worthiness, as the public's interest in verifying the claim is limited. Finally, (c) exhibits the lowest check-worthiness, as there is no factual claim in the statement. By harnessing the concept of check-worthiness, a pathway emerges for comprehensive investigations into how factual claims of varying importance influence and reflect people's behaviors on social media platforms such as Twitter.

Individuals' behaviors on social media mainly consist of posting, commenting, sharing, liking, and following. They make up the communication and information diffusion on social media. Hence, unsurprisingly many studies explored factors that influence these behaviors. For example, Comarela et al. [47] found several factors influencing user response or retweet probability, including previous responses to the same user, the user's posting rate, age, and tweet content. Firdaus et al. [48] discovered that a user's emotion towards a topic is a useful feature in modeling their retweet decision. Hopcroft et al. [49] observed that geographic distance, common friends, social status overlap, and interactions between two users (e.g., retweeting and replying) are correlated with two-way following relationships. Although a lot of studies explored the factors influencing posting and following behaviors on social media, none of them has looked into the impact of check-worthiness. A few slightly related studies mainly focused on the strategies and effectiveness of fact-checks [50, 33, 51] rather than people's behavioral tendencies toward factual claims.

Considering our limited understanding of the impact of check-worthiness on social media behaviors, it is meaningful to give a thorough investigation into it. Therefore, in this study, we conduct a range of experiments aimed at unveiling the underlying connections between factual claims and the conduct of users on social media. Specifically, these experiments leverage the concept of check-worthiness to determine whether individuals tend to follow and endorse others who exhibit comparable check-worthiness levels in their posts, whether individuals with similar behavioral tendencies toward check-worthiness exhibit some common attributes such as popularity and interests, and so on. We performed statistical analyses such as correlation analysis and hypothesis testing on numerous tweets to address those questions, thereby unveiling the impact of check-worthiness on individuals' tweeting, liking, and following behaviors.

This study acquired 3 datasets from Twitter, comprising approximately 48.5 million tweets and 15,000 users. These datasets encompass a range of general topic domains including literature, arts, religion, and politics.

Our experiments on these datasets identified several pronounced results—(1) People do express different behavioral tendencies toward factual claims; (2) People with backgrounds and interests in media, politics, and technology tend to engage more frequently with factual claims, while people related to literature, arts, and religion generate fewer factual claims; (3) Individuals tend to share and like posts with similar check-worthiness levels as their own posts; (4) In instances where individuals followed each other, there is a higher likelihood of similar behavioral tendencies towards factual claims when compared to one-way following relationships.

In summary, our contributions can be delineated as follows:

- Pioneering investigation into the interplay between check-worthiness and user behaviors within social media.
- Provision of an expansive dataset comprising around 48.5 million tweets and $15,000$ users, encompassing different types of tweets and domains.
- The outcomes of our experiments yield several noteworthy revelations that have the potential to stimulate diverse studies such as behavioral modeling and recommendation systems, particularly those pertaining to population behavior patterns in the context of social media platforms.

3.2 Related Work

Claim detection—identifying claims warranting fact-checking—is a task in the workflow of fact-checking in which check-worthiness is conceptualized. It can be viewed as a binary classification task and further as a ranking task on the claims. Numerous works have been dedicated to various modeling methodologies and their

evaluations for claim detection [42, 43, 44, 45, 46]. Furthermore, researchers have used it for various application contexts, including automated check-worthy claims collection, factual claims visualization, fake news detection, and so on [8, 52, 53].

Research on identifying factors that influence posting, sharing, liking, and following behaviors in social media is directly pertinent to our study. There are many existing works on this topic. Comarela et al. [47] conducted an extensive characterization of a large Twitter dataset which includes all users, social relations, and messages posted from the beginning of the platform up to August 2009. They identified several factors that influence user responses or retweet probability, such as previous interactions with the same tweeter, the tweeter's posting frequency and age, and entities in tweets such as hashtags and mentions. Firdaus et al. [54] uncovered that a user's emotional state can influence their retweeting behavior. They demonstrated this by constructing a retweet prediction framework based on an emotion detection model and conducting experiments using the Stanford Twitter Sentiment dataset [55] and the Obama–McCain Debate dataset [56]. Hopcroft et al. [49] identified that geographic distance, common friends, social status overlap, and the interactions between two users are correlated with two-way following relationships.

While there exists an abundance of research that directly but separately addresses check-worthiness and human behaviors on social media, to the best of our knowledge, no existing work has linked them together. There are works with a focus on analyzing the relationship between fact-checks and audience behaviors. For example, Kim et al. [50] analyzed 914 news articles with fact-checks in South Korea. They found that news articles triggered more audience comments when they mentioned the importance of fact-checking the claim under scrutiny and conveyed negative content. Clayton et al. [33] evaluated the effectiveness of strategies for designing fact-checks by conducting experiments among 2,994 participants recruited from Amazon Mechanical

Turk. Their experiments found the "Rated False" tag is more effective than the "Disputed" tag and the effect of a general warning is small compared to these two tags. Park et al. [51] discovered the unexpected and diminished effect of fact-checking due to cognitive biases. They found that claims labeled "Lack of Evidence" were often treated as false, revealing an uncertainty-aversion bias, and users who initially disapproved of a claim were less likely to change their views when presented with opposite fact-checking labels, indicating a disapproval bias. All these studies concentrated on scrutinizing the connections between fact-checks and their audiences' behaviors, primarily with the goal of understanding people's responses to truthfulness of claims. In contrast, our study is centered on discerning the correlation between check-worthiness of claims and people's behaviors.

3.3 Research Questions

Our research focuses on investigating how check-worthiness of claims impacts or reflects social media users' behaviors. There are many different angles and means to tackle this topic such as analyzing users' posting and retweeting behaviors. Regardless of the approaches, the essence of this topic lies in determining whether check-worthiness can serve as an indicator to capture people's behavioral patterns in common and to differentiate among groups of individuals. In our study, we examine this subject by observing the tweeting, following, and liking behaviors among different groups of Twitter users since those behaviors make up most of their activities on Twitter.

To investigate the impact of check-worthiness, our initial inquiry revolves around determining whether individuals exhibit varying behavioral tendencies when confronted with factual claims of differing check-worthiness levels (i.e., showing higher engagement with claims of low/high check-worthiness). If such disparity arises, we

then dig into the underlying rationales and figure out whether this disparity has an influence on individuals' behaviors or reflects unique characteristics of the respective populations. To unravel these unknowns, we introduce the following specific research questions for our study.

Q1. Do people exhibit different behavioral tendencies toward check-worthiness?

Q2. What factors and commonalities within the population might account for such behavioral tendencies?

Q3. Do people maintain similar check-worthiness levels between the tweets they post and those they favor?

Q4. Do people tend to follow others who exhibit similar behavioral tendencies toward check-worthiness?

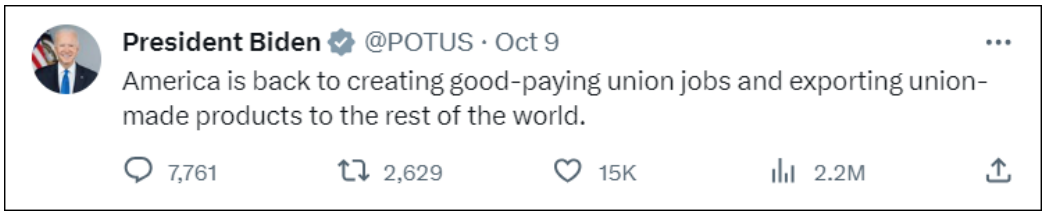
3.4 Datasets

The smallest research unit in this study is a tweet, which is a social media post published by a Twitter user. For the convenience of wording, we categorize tweets into 3 types: *original-tweet*—a tweet that is initially created by a Twitter user; *retweet*—a tweet that is reposted by a Twitter user from an original-tweet; and *liked-tweet*—a tweet that is liked by a Twitter user. Figure 3.2 shows examples of them.

We collected 3 datasets to support our study, which are accessible at *Zenodo*. [1] Due to Twitter's content redistribution policy, the datasets only include tweet IDs and user IDs instead of tweet content and user profile details. Table 3.1 provides basic statistics about these datasets. Detailed descriptions of the datasets are as follows.

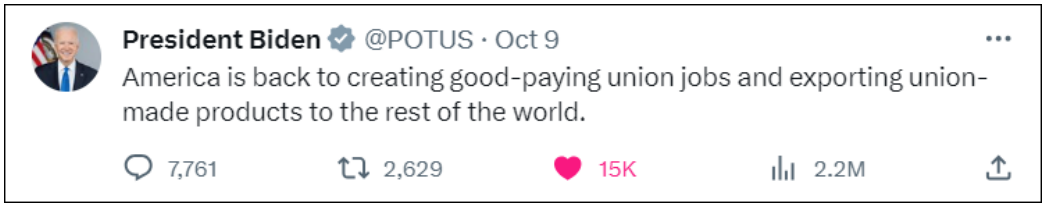
Randomly Sampled Users Dataset (RSU) This dataset consists of $11,173$ users collected through Twitter's APIs. We collected $10,000$ random English tweets in

[1] https://zenodo.org/records/11081026



(a) Original-tweet

(b) Retweet



(c) Liked-tweet

Figure 3.2: Types of tweets

Dataset	Tweets Count	Users Count	Collection Date	Domain
RSU	40,405,150	11,173	02/2023	Random
POL	8,153,745	3,784	05/2023	Politics
HUM	341,285	498	01/2024	Literature, Art, Philosophy, and Religion

Table 3.1: Datasets statistics

February 2023 using Twitter's Volume Stream API. The tweets were posted by around $3,000$ users. For each user, we collected up to 100 of its most recent followees using Twitter's Following API. Through the Timeline and Liking APIs, for each user, we collected their most recent tweets (up to $3,200$ tweets due to Twitter's limit) and liked-tweets (up to $3,200$ too). We then filtered out users that have insufficient tweets (less than 100 original-tweets or less than 80 retweets/liked-tweets) to ensure that the

sample sizes are statistically significant in our analyses. Finally, we have 11,173 users along with 40,405,150 tweets. To prevent potential sampling biases in the collected data, we randomly selected 50 users and examined their profiles and their most recent 30 tweets to detect any biases (e.g., concentration in their backgrounds and topics). The results showed no significant concentration in the users' backgrounds or topics.

Politics Dataset (POL) This dataset contains all tweets from selected U.S. news media and U.S. politicians including *Senators*, *House Members*, *US Governors*, *US Secretaries of State*, *US Cabinet*, and *US Election Officials* at collection time. We used Twitter's Timeline API to collect the most recent tweets (up to 3,200) of the target accounts. The dataset was collected in May 2023, with 8,153,745 tweets and 3,784 Twitter accounts.

Humanities Dataset (HUM) This dataset contains 341,285 tweets and 498 Twitter accounts from selected Twitter lists including *Writers*, *Christianity*, *Artists*, *Buddhism*, *Musicians*, and *Philosophers*. We use Twitter's List and Timeline APIs to collect the accounts and their most recent tweets (up to 1,000). The dataset was collected in January 2024.

3.5 Methodology

Each tweet in our datasets is associated with a corresponding check-worthiness score to indicate how check-worthy it is. We employed the ClaimBuster [42] API [2] to obtain check-worthiness scores. Given a tweet, the API returns a score ranging from 0 to 1, corresponding to how likely the tweet contains a check-worthy factual claim.

[2]https://idir.uta.edu/claimbuster/api

ClaimBuster has been used by researchers and fact-checkers in various contexts. For instance, the Duke Reporters' Lab [3] used ClaimBuster to create daily email alerts to professional fact-checkers with the most check-worthy claims from TV program transcripts and social media. These alerts have led to at least 33 claims featured in 30 different articles by fact-checking outlets, including one from The Washington Post that was discussed in a news report [57]. It has been applied in real-time for the live coverage of all primary election and general election debates of the U.S. presidential elections since 2016. Post-hoc analysis of the claims checked by professional fact-checkers at *CNN*, *PolitiFact.com*, and *FactCheck.org* reveals a highly positive correlation between ClaimBuster and fact-checkers in deciding which claims to check [7].

Although ClaimBuster has been widely applied in presidential debates, political speeches, and interviews, it is worth assessing its effectiveness on tweets, which are less formal and noisier. We did not use public datasets such as CLEF CheckThat! [4] for evaluation because their data includes multimodal features (e.g., images) and does not align perfectly with our evaluation criteria. Our labels differ from theirs by 20% in a random sample of 100 tweets from their dataset. To this end, we conducted a human evaluation on a random sample of 200 tweets selected from our datasets. Each tweet was annotated by 3 annotators who labeled them as either check-worthy or non-check-worthy. All of the 3 annotators possess the concept and experience of check-worthiness evaluation as they all have contributed to factual claim detection tasks. The final label of each tweet was decided by majority vote. We used a check-worthiness score threshold of 0.5 to classify the tweets: If a tweet received a ClaimBuster score above 0.5, it would be classified as check-worthy; otherwise, non-check-worthy. This simple

[3] http://reporterslab.org/tech-and-check

[4] https://checkthat.gitlab.io/clef2024/task1

classifier has an accuracy of 0.84, indicating that ClaimBuster is effective in identifying check-worthy tweets and thus it can be used as a reliable tool for analyses in our study.

Our study frequently utilizes correlation analysis and hypothesis testing in the experiments as they are simple and useful tools for identifying and verifying underlying connections between variables. In this study, we use a scatter plot to visualize the relationship between two variables and use the Pearson correlation coefficient [58] to measure the direction and strength of a linear relationship between two variables.

Hypothesis testing is widely used in verifying statistical conjectures by examining data samples. We use it to validate our presumption about individuals' behavioral tendencies toward check-worthiness. We refer to H_0 as the null hypothesis, for which we test whether to accept it. If we reject it, we will accept the alternative hypothesis H_a. In this study, we primarily use hypothesis testing to assess the equality of check-worthiness distributions across thousands of user sets, aiming to ascertain the similarities or differences between individuals' tweeting behaviors. When conducting the same hypothesis test many times using different data, one may observe some statistically significant results just by chance, even if there is no true effect. As we are performing some hypothesis tests thousands of times in the experiments in Section 3.6, the test results might contain many false positives by chance. Therefore, the false discovery rate (FDR) using the Benjamini-Hochberg procedure [59] was applied to control false significant results by adjusting the p-values.

Generally speaking, when determining if two samples originate from the same distribution, our preference would be *Z-test* or *T-test* in instances where we have equally sized samples and can make the assumption that the underlying populations adhere to normal distributions with known variances. Nonetheless, the check-worthiness of a Twitter user's posts hardly conforms to a normal distribution. We substantiated this claim by performing Shapiro–Wilk tests [60] on the randomly sam-

pled users dataset RSU (Section 3.4). Shapiro–Wilk test is one of the most popular hypothesis tests for examining how close the sample data fit to a normal distribution by ordering and standardizing the sample. Given each user in the RSU dataset, we performed Shapiro–Wilk test with significance level $\alpha = 0.05$ on the check-worthiness scores of the user's original-tweets, retweets, and liked-tweets, respectively. The result, as displayed in Table 3.2, shows that only a few of the null hypotheses were accepted across all users and tweet types. This suggests a very low probability that the check-worthiness scores of a user's original-tweets, retweets, or liked-tweets follow a normal distribution.

H_0 ($\alpha = 0.05$)	**Accept**	**Reject**
The check-worthiness scores of a user's original-tweets are normally distributed	37	11136
The check-worthiness scores of a user's retweets are normally distributed	259	10914
The check-worthiness scores of a user's liked-tweets are normally distributed	58	11115

Table 3.2: Normality test on check-worthiness distributions

Given that it is highly unlikely the check-worthiness scores follow normal distributions, *Z-test* and *T-test* become less applicable. Hence, we selected two non-parametric tests that are applicable under less rigorous conditions—Brunner Munzel test [61] and Kolmogorov-Smirnov test [62]. Both tests possess the capability to assess the stochastic equality of two random variables—whether one is "larger" than another—without rigorous assumptions such as identical distribution type and equal variances. The Kolmogorov-Smirnov test is more strict since it tests whether two samples are from the same distribution, while the Brunner Munzel test only exam-

ines the stochastic equality of two samples. We articulate the formal null hypotheses of these two tests as follows.

- **H_0 of Brunner Munzel (BM) test**: For randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X.
- **H_0 of Kolmogorov-Smirnov (KS) test**: Two sets of samples are drawn from the same (but unknown) probability distribution.

3.6 Experiments

3.6.1 Q1: Individuals' Behavioral Tendencies Toward Check-Worthiness

The very first question we want to answer is whether people exhibit different behavioral tendencies toward check-worthiness. The RSU dataset contains a large number of random users and their corresponding tweets, making it a suitable dataset for investigating this query.

The most straightforward way of checking an individual's behavioral tendency toward check-worthiness is the overall check-worthiness of their posts. Hence, for each user in the RSU dataset, we computed the median check-worthiness score of their tweets, denoted as *individual check-worthiness*. We chose the median because check-worthiness scores are typically not normally distributed and tend to be skewed. We present in a histogram (Figure 3.3) the distribution of individual check-worthiness of all users in the RSU dataset. It shows that, although individual check-worthiness mostly concentrates between 0.3 and 0.4, there are people who exhibit a particular tendency toward higher or lower check-worthiness. That motivates us to explore more about the underlying rationales and behavioral consequences of those preferences.

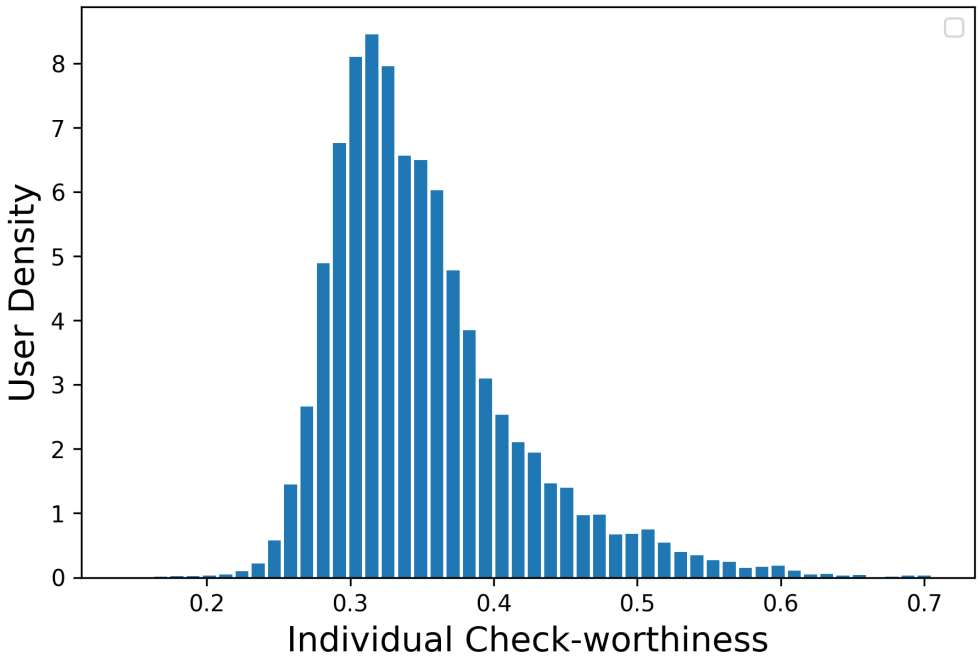


Figure 3.3: Individual check-worthiness distribution

3.6.2 Q2: Causes of Different Behavioral Tendencies Toward Check-Worthiness

Knowing the difference between people's behavioral tendencies toward check-worthiness prompts us to speculate whether there are some common attributes correlated with those preferences. To investigate this question, we conducted correlation analyses on various features of users in the RSU dataset.

First of all, we analyzed the numeric features—posts count, favorites count, followers count, followees count, listed count (number of lists containing the user), and media count (number of posts containing images or videos). These features primarily reflect a user's popularity and activity level. We performed a univariate correlation analysis by calculating the Pearson correlation coefficients between individual check-worthiness and log-transformed feature values. The results, as Figure 3.4 shows, are all weak correlations along with most p-values less than 0.005. In addition, a

multivariate regression analysis on these features yielded both tiny coefficients and an R^2 value of 0.25, indicating a weak correlation between individual check-worthiness and these features. Based on the results, we cannot conclude that features related to popularity and activity level are indicators of individuals' behavioral tendencies toward check-worthiness.

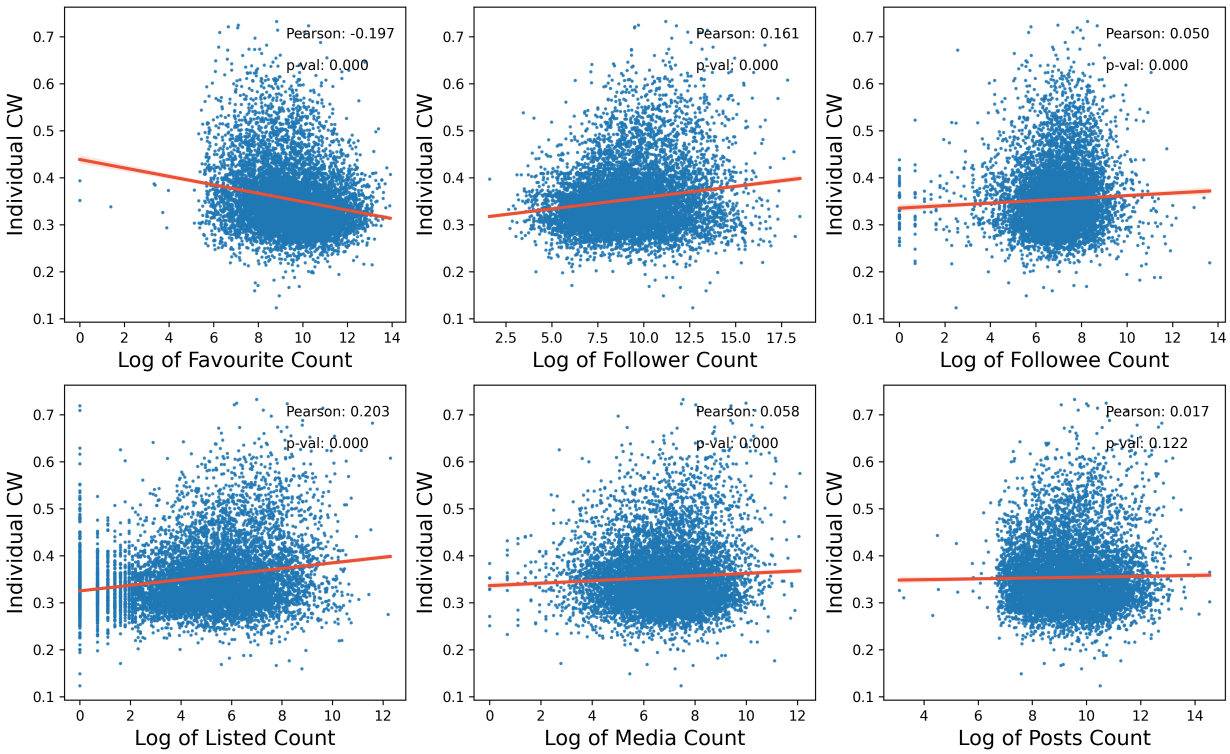


Figure 3.4: Correlation between individual check-worthiness and popularity/activity features

Besides those numerical features representing popularity and activity levels, there are other more complex features that could affect/indicate the tendencies. Such

features may include occupation, political spectrum, and educational background. Since these features are either hidden or challenging to identify using automatic methods, we decided to conduct a content analysis to obtain some insights.

Firstly, among the $11,173$ users in the RSU dataset, we selected all users with individual check-worthiness scores less than 0.25 or greater than 0.55 as they encompass two tails of the individual check-worthiness distribution, and thereby represent two small groups with weak and strong behavioral tendencies toward check-worthiness. The group denoted as U_0 comprises 146 users with low individual check-worthiness, while the group denoted as U_1 consists of 169 users with high individual check-worthiness. For both groups, we gathered all the user profile descriptions and tweets from the users, and conducted word frequency analysis. More specifically, for all the user profile descriptions and tweets respectively, we tokenized, removed stopwords, lemmatized, and counted word frequencies. The results from U_0 and U_1 are quite different, as Table 3.3 shows. The top frequent words in tweets from U_0 are general and irrelevant to specific people/events/affairs (e.g., love, life, like, god, good), while the top frequent words in tweets from U_1 are more concrete and highly related to trending topics/events (e.g., russian, ukraine, cannabis). The analysis of the user profile descriptions further enhances this observation. The top frequent words in user profile descriptions from U_0 are more related to literature, life, entertainment, and religion, while the top frequent words in user profile descriptions from U_1 are more related to journalism, politics, and technology.

The results from the word frequency analyses appear to suggest that individuals' professions, backgrounds, and interests are potentially related to their behavioral tendencies toward check-worthiness. To confirm this conjecture, we randomly selected 100 users from U_0 and U_1 respectively, and then we annotated each user account based on their backgrounds and interests. The results, as shown in Table 3.4, reveal that

U_0's Tweets	U_1's Tweets	U_0's Profiles	U_1's Profiles
people (13112)	new (9591)	author (15)	news (23)
love (12462)	russian (6841)	podcast (7)	reporter (12)
life (12419)	ukraine (5135)	com (7)	newsletter (9)
one (10759)	people (4152)	book (7)	com (9)
like (9956)	energy (3906)	producer (6)	world (8)
god (9138)	report (3667)	life (5)	public (7)
us (8654)	said (3576)	people (5)	tech (7)
time (8249)	russia (3485)	views (5)	government (7)
get (7506)	cannabis (3215)	buddhist (5)	research (6)
good (7368)	us (3135)	writer (5)	policy (6)

Table 3.3: Frequent words in tweets and profiles from U_0 and U_1

U_0's BGs	U_1's BGs	U_0's Interests	U_1's Interests
unknown (36)	media (33)	ideology (40)	politics (25)
writer (19)	reporter (13)	daily life (39)	general news (14)
influencer (12)	research (9)	religion (7)	public interest (12)
pastor (3)	politician (6)	entertainment (3)	tech/science (12)
speaker (3)	analyst (5)	photography (2)	climate (9)
singer (2)	journalist (4)	writing (2)	energy (8)
photographer (2)	unknown (4)	general (2)	security (7)
consultant (2)	writer (4)		business (6)
student (2)	advocate (4)		war (3)
teacher (2)	editor (3)		economics (2)

Table 3.4: Top-ranked backgrounds and interests in U_0 and U_1

a majority of users in U_0 lack explicit backgrounds, though a considerable portion comprises writers and influencers. Their primary interests lie in sharing their ideologies and daily lives. On the other hand, users in U_1 are prominently associated with the media, with over half of the selected 100 users actively doing media-related jobs. Additionally, users with backgrounds in research and politics are also notably present. The dominant interests within U_1 encompass politics, general news, public interests, and technology/science, validating our initial conjecture.

To further solidify this conclusion, we also compared the individual check-worthiness distributions of three specific groups of Twitter users using the afore-

mentioned datasets. The first group contains all users from the HUM dataset, which encompasses individuals related to humanities such as literature, arts and religion. The second group consists of all users from the POL dataset, which represents individuals related to politics and journalism. The third group consists of all users from the RSU dataset which are randomly sampled user accounts. Figure 3.5 depicts the individual check-worthiness distributions of these three groups of users. The figure shows a left-skewed distribution for the HUM dataset, a right-skewed distribution for the POL dataset, with the RSU in the middle. This finding suggests that users in the HUM dataset generally possess lower individual check-worthiness, whereas those in POL tend to exhibit higher levels of check-worthiness.

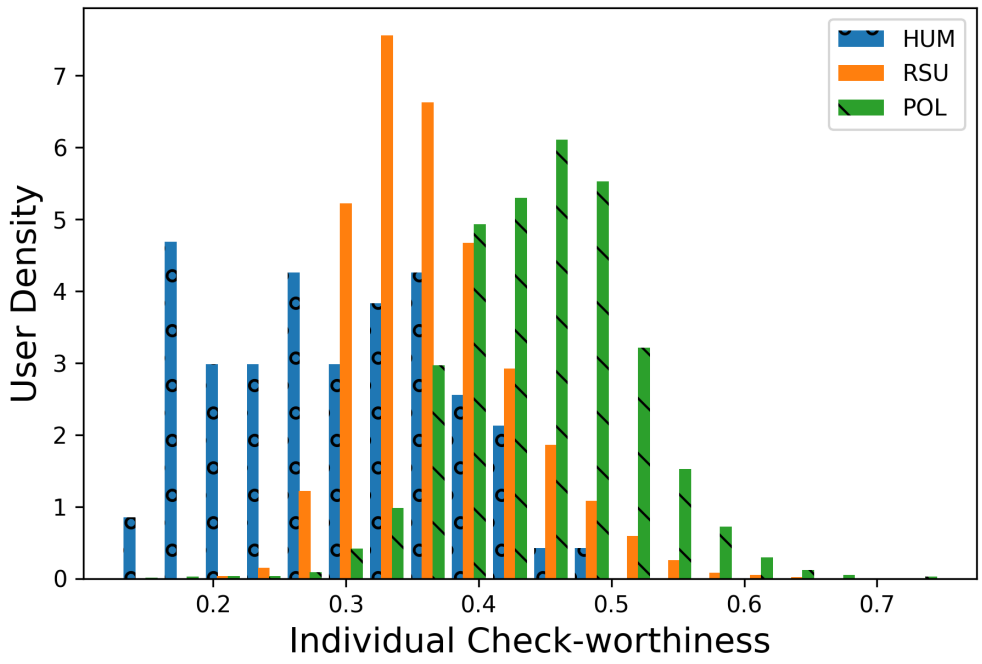


Figure 3.5: Individual check-worthiness distributions for HUM, RSU, and POL

3.6.3 Q3: Impact of Check-Worthiness on Tweeting Behaviors

With the conclusion that people express different behavioral tendencies toward check-worthiness, one may ask how well it aligns with people's tweeting behaviors. More specifically, it would be useful to know whether people tend to share and like posts with similar check-worthiness as their own posts. To answer this question, we conducted experiments on the RSU dataset.

We define O, R, and L as random variables of the check-worthiness of a randomly picked original-tweet, retweet, and liked-tweet from a given user. Moreover, given the dataset, we define X and P as random variables of the check-worthiness of a random tweet and a random popular tweet (liked or retweeted by anyone) from that dataset. Hence, with the RSU dataset, we have 4 hypotheses defined as follows to test the stochastic equality between O and other random variables:

- **Hyp1**
$$\begin{cases} H_0: & P(O > R) = P(O < R) \\ H_a: & P(O > R) \neq P(O < R) \end{cases}$$
- **Hyp2**
$$\begin{cases} H_0: & P(O > L) = P(O < L) \\ H_a: & P(O > L) \neq P(O < L) \end{cases}$$
- **Hyp3**
$$\begin{cases} H_0: & P(O > P) = P(O < P) \\ H_a: & P(O > P) \neq P(O < P) \end{cases}$$
- **Hyp4**
$$\begin{cases} H_0: & P(O > X) = P(O < X) \\ H_a: & P(O > X) \neq P(O < X) \end{cases}$$

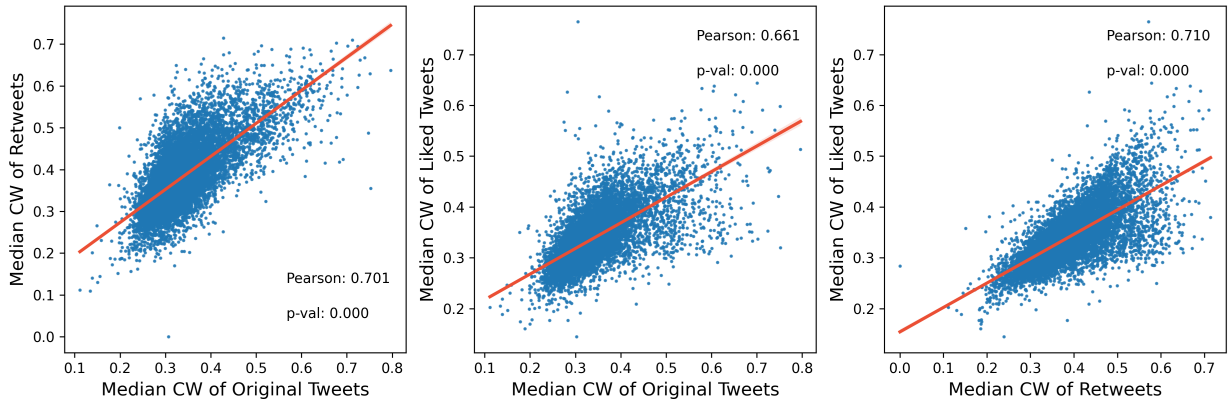
For each user in the RSU dataset, we performed Brunner Munzel (BM) test and Kolmogorov-Smirnov (KS) test on Hyp1-4, as explained in Section 3.5. Table 3.5 shows the results of the acceptance for those hypotheses with alpha (significance

level) equal to 0.05. We can see that the acceptance rates of Hyp1-2 are greater than that of Hyp3-4 for all the tests, which means the check-worthiness distribution of a user's original-tweets is more likely to have the same shape as the check-worthiness distributions of the same user's retweets and liked-tweets, in comparison to random and popular tweets from arbitrary users.

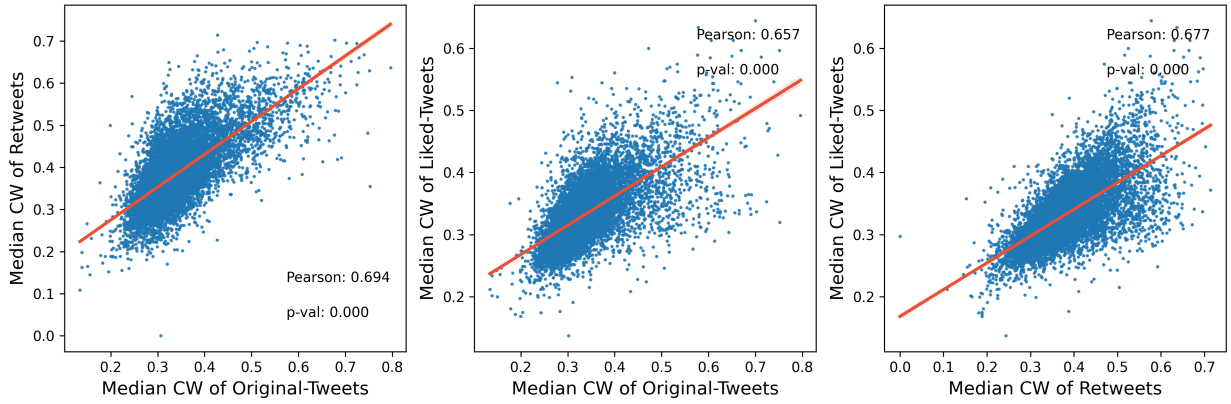
Test	**Hyp1**	**Hyp2**	**Hyp3**	**Hyp4**
BM Test	1797/16.1%	2896/25.9%	939/8.4%	1013/9.1%
KS Test	1126/10.1%	1831/16.4%	248/2.2%	284/2.5%

Table 3.5: Acceptances of Hyp1-4

In addition, we also performed a correlation analysis on the median check-worthiness of original-tweets, retweets, and liked-tweets for all the users in the RSU dataset. As Figure 3.6(a) shows, where each point represents a user account, there exist strong correlations between the median check-worthiness scores of the users' original-tweets, retweets, and liked-tweets. To reduce the effects of content redundancy (such as when a user retweets or likes their own original-tweet, or when a retweet is also liked by the retweeter), we computed the overlap ratios among these three types of tweets for each user. The result shows that, on average, merely 1.3% of original-tweets are additionally retweeted by their authors, and less than 1% of original-tweets are also liked by the authors. Additionally, about 16% of retweets are also liked by the retweeters. To eliminate the impact of overlapping tweets, for each tweet from a given user that appeared in more than one category (i.e., original-tweets, retweets, and liked-tweets), we retained it in only one category, following the priority order original-tweets $>$ retweets $>$ liked-tweets. After the cleaning, we conducted the correlation analysis again, and the results, as shown in Figure 3.6(b), remained largely unchanged. Therefore, we are able to conclude that people overall have the



(a) Without removal of overlapping tweets



(b) With removal of overlapping tweets

Figure 3.6: Correlation in median CW among three types of tweets

behavioral tendency to post, share, and favor tweets with similar check-worthiness levels.

3.6.4 Q4: Impact of Check-Worthiness on Following Behaviors

Besides tweeting activities, another important activity on social media is following, which influences a large portion of the information a user receives. Therefore, it is natural and crucial to find out whether people tend to follow others with similar ten-

dencies toward check-worthiness. More specifically, we want to examine whether the check-worthiness distribution of a user's tweets is more similar to that of its followers than other users.

We define U, V, F, and X as random variables of the check-worthiness of a randomly picked tweet from a given user, one of its followers, one of its friends (being both follower and followee), and a random user respectively. Here we have the hypotheses defined as follows to test the stochastic equality between U and other random variables:

- **Hyp5** $\begin{cases} H_0: & P(U > V) = P(U < V) \\ H_a: & P(U > V) \neq P(U < V) \end{cases}$
- **Hyp6** $\begin{cases} H_0: & P(U > F) = P(U < F) \\ H_a: & P(U > F) \neq P(U < F) \end{cases}$
- **Hyp7** $\begin{cases} H_0: & P(U > X) = P(U < X) \\ H_a: & P(U > X) \neq P(U < X) \end{cases}$

In the RSU dataset, we have $10,402$ (follower, followee) pairs and 351 friend pairs, with a total of $9,124$ distinct accounts involved. Table 3.6 shows the results of the acceptance for hypotheses Hyp5-7 with alpha (significance level) equal to 0.05. We can see that the acceptance rates of Hyp5 are greater than Hyp7, meaning the check-worthiness distribution of a user's tweets is more likely to have the same shape as the check-worthiness distribution of its followers' tweets compared with a random user's tweets. However, this likelihood is not strong since the acceptance rates of Hyp5 do not exceed Hyp7 by a lot. A more substantial result comes from the acceptance rates of Hyp6, which are much higher. This indicates a higher likelihood of check-

worthiness similarity between tweets from a pair of users in a two-way following relationship than in a one-way following relationship.

Test	**Hyp5**	**Hyp6**	**Hyp7**
BM Test	1043/10%	59/16.9%	696/7.6%
KS Test	335/3.2%	50/14.3%	130/1.4%

Table 3.6: Acceptances of Hyp5-7

To further verify this conclusion, we again performed a correlation analysis on the individual check-worthiness of users and their friends. As Figure 3.7 shows, there exists a weak correlation between the individual check-worthiness of followers and followees. However, the correlation becomes stronger when we compare the individual check-worthiness of users with a two-way following relationship. Similar to what we discussed in Section 3.6.3, our calculation shows that the average ratio of tweet overlap between each pair is less than 1%. This means the possibility of the result being influenced by overlapping tweets among friends is low. Therefore, the result is solid and aligns with our conjecture.

3.7 Limitation

While we have examined certain social media behaviors and identified several patterns, a few questions about the observations remain unanswered. For example, since most users may not consciously choose high or low check-worthy content when posting, sharing and liking tweets, the observations cannot be interpreted as definitive indicators of users' behavioral patterns. However, our statistics reveal that a significant portion of users demonstrate consistent behavioral patterns associated with check-worthiness, suggesting a possible subconscious adherence to such patterns. Undoubtedly, a more nuanced investigation is warranted, particularly concerning where

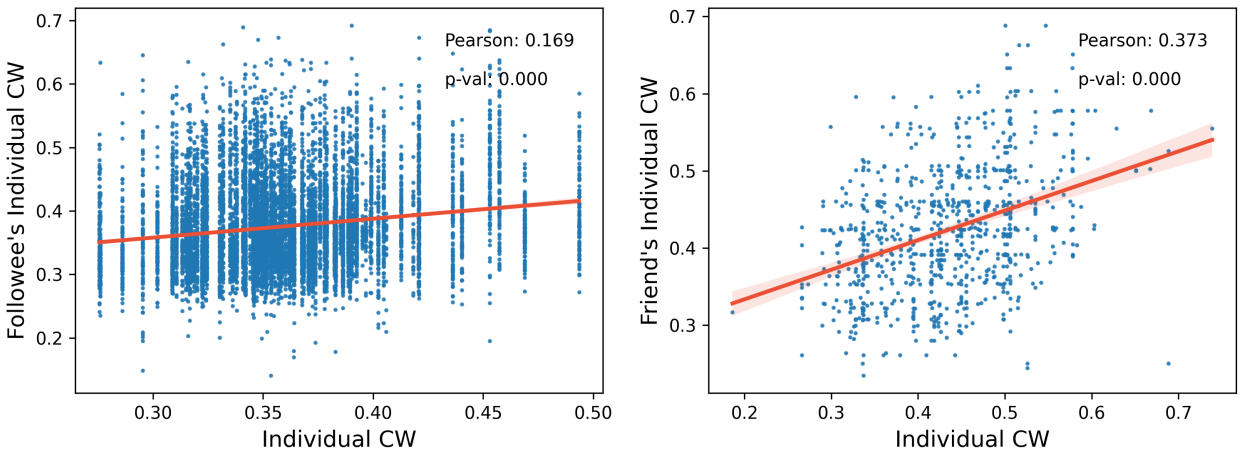


Figure 3.7: Correlation between following parties' individual check-worthiness

the observed patterns come from. The inherent value of our study lies not only in the answers it provides but also in the thought-provoking questions it raises. This study possesses the potential to not only address current gaps in knowledge but also to act as a catalyst, possibly inaugurating a new line of research endeavors.

3.8 Conclusion

This study identified the existence of difference between individuals' behavioral tendencies toward factual claims. In particular, the population from domains such as politics, general news, and technology is more likely to engage with check-worthy content compared to those associated with arts, literature, and religions. Through a set of experiments, the research has established a strong correlation between these tendencies and users' posting, sharing, and liking behaviors, indicating a conspicuous pattern to engage with content of similar check-worthiness. Furthermore, the findings emphasize the heightened efficacy of two-way following relationships in reflecting shared preferences towards factual claims.

The concept of check-worthiness emerges as a potent tool for understanding human behaviors within the realm of social media. Our results not only provide valuable insights into the impact and adaptability of check-worthiness but also lay the groundwork for future investigations to delve deeper into the various dimensions of its influence on social media behaviors and its potential applications across diverse domains.

CHAPTER 4

WILDFIRE: A SOCIAL SENSING PLATFORM FOR LAYPERSON

4.1 Introduction

As mentioned in Chapter 1, social sensing offers a means to observe and interpret phenomena and discover insights about our society, providing opportunities beyond traditional methods. It can scale up to millions of users within just a few days, in contrast to traditional surveys, and can capture real-time changes in public opinion, setting it apart from traditional polls [63]. More specifically, online data production has some similarities to traditional data-collecting techniques like surveys and structured observations, and it also has some distinctive characteristics and new features [64]. In contrast to surveys, which only allow for retroactive memory of actions or feelings, social media platforms allow for widespread, in-the-moment observation of human interaction and personal expression. Instead of the years generally needed for survey-based data gathering, social scientists may analyze social media data and start presenting conclusions within a couple of months (or sooner) with the right infrastructure. Social media data also exhibit several traits common to ethnographic or observational research. For example, data from social media allow researchers to get unprompted and voluntary behaviors [65]. One may even argue that these data offer a more accurate representation of typical social interactions than some routine social experiments.

Nevertheless, in the past, it was not popular to use social media data for social science studies because of the many discrepancies between social media data and data gathered using conventional surveys. For instance, when utilizing conventional

surveys, researchers may have few respondents but can control what data the respondents offer [63]. In these circumstances, respondents offer researchers information that is valuable, but the small sample size may not create enough diversity to fully examine phenomena that are encountered less often. On the contrary, social media data provide huge volume and diversity but it is hard to control the content of data. Nowadays, with the development of the digital world, social sensing has become an important way to conduct social studies. It can offer unique insights from data that only exist in social media, such as "likes". Furthermore, technological advancements in deep learning and natural language processing have greatly amplified the benefits of social sensing via sophisticated, large-scale analysis of social media data [66, 67].

Many of those who need to conduct social sensing in the real world are non-technical users, but social sensing requires complex skills in coding, data modeling, data processing, and analytics. Social media data can be abstracted as a graph with posts, users, and their relationships. It is non-trivial to retrieve the data from platforms (e.g., through various Twitter APIs, [1] using search and query conditions), to parse the data which could be in formats such as JSON and may contain both texts and multimedia, to store them in accordance with well-designed schema, and to serve the stored data based on search and filter conditions. All these are particularly challenging given the large volume and velocity of social media. Given practical constraints such as the access rate limits in Twitter APIs, one often needs to devise sophisticated algorithmic approaches to prioritizing which part of the graph to retrieve before others. Once the data are in place, various types of content analysis (e.g., text classification, sentiment analysis, stance detection), network analysis (e.g., ranking users based on importance), and data visualization are performed. These analytics may need to be repeated periodically given the constantly updated data.

[1] https://developer.twitter.com/docs/twitter-api

This study introduces Wildfire, a novel social sensing platform focusing on Twitter data. It enables anyone with access to Twitter APIs to conduct social sensing tasks on large volumes of data, without writing a single line of code. Our contributions are:

- Wildfire offers a flexible data collection mechanism, harnessed through Twitter's APIs, and a heuristic graph exploration approach. This method expands the initial seed collection of tweets and accounts by employing a ranking function to iteratively identify target accounts and tweets via 'following' relationships between accounts. The ranking function is guided by a set of weighted classifiers, including those developed in-house for tasks such as detecting check-worthiness of factual statements [68], as well as classification models using ChatGPT [2] and from HuggingFace. [3]
- Wildfire uses a single graph to store all Twitter accounts and tweets collected for multiple tasks. Meanwhile, collected data are annotated to discern the portion of the graph associated with each specific task. Using this approach Wildfire avoids redundant collection and storage of data.
- Wildfire offers a range of analytic tools (e.g., text classification, topic generation and entity recognition) which are crucial for accomplishing common tasks in social sensing such as finding trending topics and events.
- Wildfire integrates all data collection and analytics components into a single interface, enabling users to perform social sensing tasks without the need for coding.

4.2 Related Work

There exist many different social sensing tools. Open-source tweet collection tools such as [69, 70, 71] simply utilize Twitter APIs and primarily offer basic data

[2]https://chat.openai.com/

[3]https://huggingface.co/models

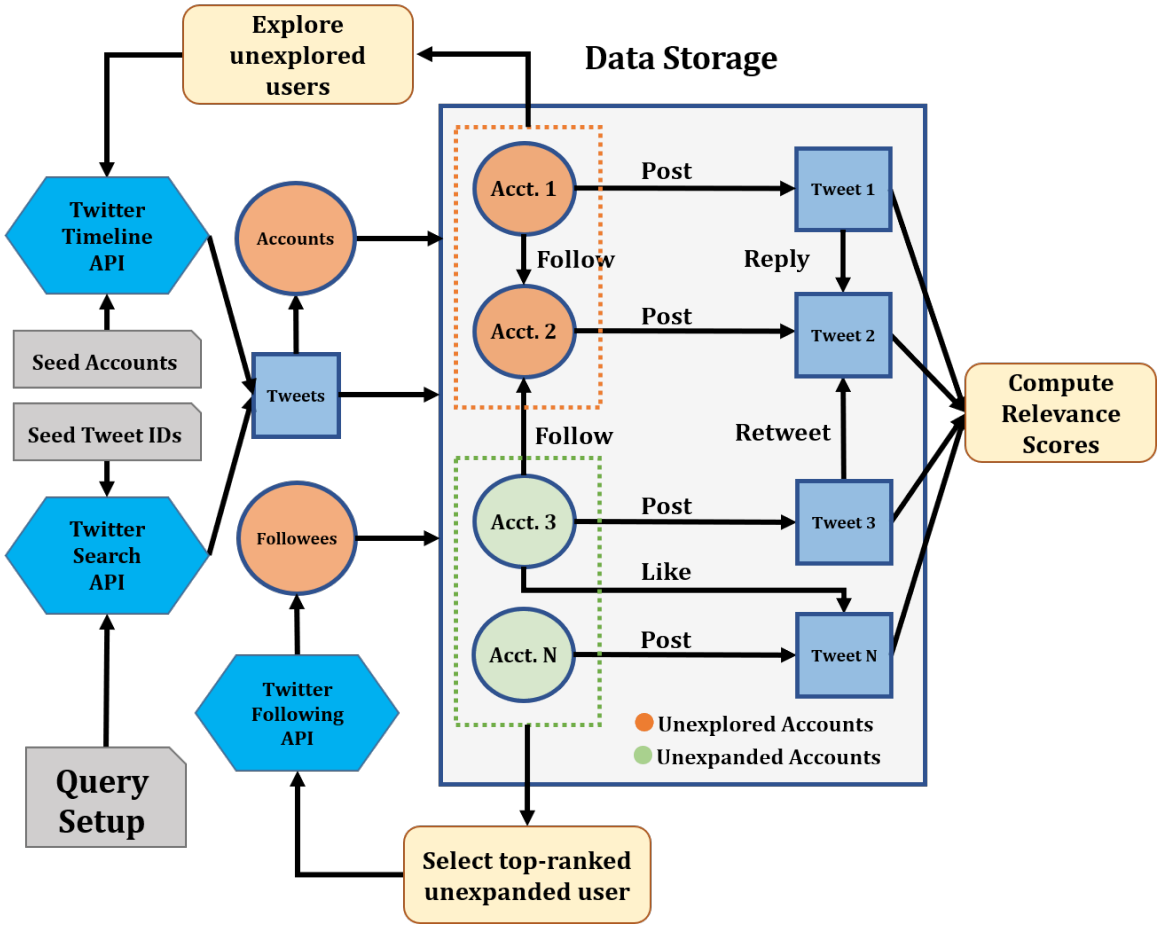


Figure 4.1: Data collection architecture

statistics such as hashtags and mention frequencies. Commercial social sensing tools (e.g., Brandwatch, [4] Digimind, [5] Hootsuite, [6] and Synthesio [7]) excel in comprehensive data analytics, but they mainly focus on business analysis and lack the capabilities to customize data collection from the perspective of academic research. For example, they do not support collecting data through graph exploration methods such as community detection algorithms. They also do not provide users the freedom to leverage

[4]https://brandwatch.com
[5]https://digimind.com
[6]https://hootsuite.com
[7]https://synthesio.com

a variety of cutting-edge machine learning models according to users' needs. Some academic studies have proposed data collection tools or methods [72, 73, 74]; however, these are primarily scrapers, frameworks, and code scripts designed for developers to integrate into their data collection pipelines.

4.3 System Design

Figure 4.1 shows the data collection architecture of Wildfire. The collected tweets and Twitter accounts conceptually form a bipartite graph. The edges in the graph correspond to relationships between accounts and tweets, including Twitter accounts posting and liking tweets, accounts following each other, and tweets retweeting, quoting and replying-to other tweets. The graph is stored in a MySQL database.

To use Wildfire, a user is required to sign up and provide a Twitter bearer token.[8] Once logged in, the user can create multiple social sensing tasks. The system uses the user-provided Twitter token to collect data for their tasks. Wildfire accommodates concurrent execution of multiple tasks from the same user. Instead of populating a separate graph for each task, Wildfire stores data across all tasks in a single graph. This avoids wasteful, redundant data retrieval and storage. Particularly, actions such as retrieving timeline tweets and follower/followee lists for a specific account will only be performed once, even if the account is included in the graphs for multiple tasks. This novel and unique design allows Wildfire to share data across tasks, resulting in reduced usage of the user's token and shortened task completion time.

Given the aforementioned system design, the ensuing discussion in this section focuses on how Wildfire collects data for one task. For that, the system populates the graph with two methods that use Twitter APIs, as follows.

[8] https://developer.twitter.com/en/docs/authentication/oauth-2-0/bearer-tokens

1) The *seed collection* method uses the Search API to obtain tweets and their corresponding Twitter accounts, based on user-specified keywords or Twitter queries.[9] Alternatively, the user can choose to provide IDs of seed tweets and seed accounts to be collected.

Since Twitter limits the number of historical tweets that can be retrieved each month and returns tweets in reverse chronological order, Wildfire requires the user to specify a few parameters when creating a search task, in order to avoid rapid exhaustion of the user's monthly quota, meanwhile ensuring evenly distributed tweets across a task's search period. The parameters include *Start datetime* (s), *End datetime* (e), *Timeslot length* (t), and *Granularity* (g). Wildfire splits the search period $[s, e]$ into $\lceil \frac{e-s}{t} \rceil$ timeslots. For each timeslot, it collects at most g tweets. Figure 4.2 illustrates the idea. Note that *Start datetime* and *End datetime* can be in the future. Before collecting tweets for each timeslot, Wildfire checks whether the end of the timeslot is past. If not, it sleeps until the end of the timeslot and then uses the Search API to retrieve past tweets in reverse chronological order.

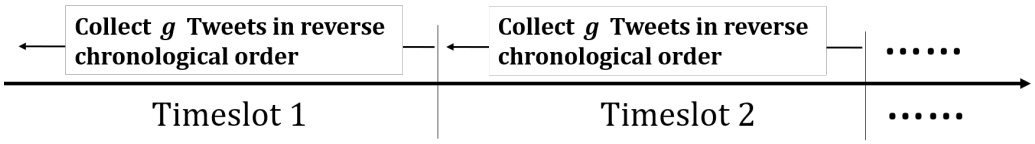


Figure 4.2: Timeslot and granularity

2) The *expansion collection* method uses the Timeline API to *explore* a Twitter account by retrieving up to 3,200 most recent tweets in its timeline and the Following API to *expand* the account by retrieving its following list (i.e., the list of accounts that this account follows). We denote accounts as *unexplored* if the system has not

[9]https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

collected their timeline tweets, and as *unexpanded* if it has collected their timeline tweets but not their following lists.

The expansion collection method operates with two concurrent, continuous processes. One process always randomly chooses an unexplored account to explore, i.e., collecting its timeline tweets. Once an account is explored, it becomes an unexpanded account. The other process employs a customized *ranking function* to select the most relevant unexpanded account to expand in hopes of finding additional accounts and tweets relevant to the given task. The use of the ranking function follows the rationale that a relevant Twitter account's followees may be relevant too since they may share common interests and beliefs and they may participate in joint or similar conversations. This expansion collection can be run concurrently with the seed collection, thereby accelerating the data collection process.

Wildfire enables users to use multiple weighted classifiers to construct their ranking functions. The classifiers aim to identify tweets exhibiting specific attributes, such as particular sentiments and political affiliations. Each individual classifier, denoted C, calculates a relevance score, denoted $C(t)$, for a given tweet t. The relevance score of a Twitter account a within the context of the classifier is the average relevance score across all N_a tweets in the account's timeline. The user has the flexibility to specify a customized ranking function by assigning a weight (w) to each of these account relevance scores from different classifiers. Users also have the option to score tweets and accounts in ascending order instead of descending order if lower values align with their interests, e.g., a user wishing to collect tweets with negative sentiment would prefer lower score values. Equation 4.1 shows how the ranking function selects the top-ranked account from the pool of unexpanded accounts (X) using k classifiers, as follows.

$$Top(X) = \underset{a \in X}{argmax} \left(\sum_{i=1}^{k} w_i * \frac{\sum_{t \in a} C_i(t)}{N_a} \right) \tag{4.1}$$

Currently, Wildfire provides 2 built-in classifiers: *ClaimBuster* [68], a deep learning model to detect tweets containing check-worthy factual claims, and *RoBERTa sentiment analyzer* [75], a multilingual model to discover the sentiment of tweets. Besides them, users can also utilize any classifier from HuggingFace by specifying the model name and the desired class (e.g., preferring negative sentiment). Wildfire employs these classifiers through HuggingFace Inference API calls, providing users with a diverse array of models to select from. Furthermore, Wildfire provides users with the option to employ ChatGPT, a chatbot based on a series of large language models [76], as a classifier by simply providing a prompt that leads to relevance scores from ChatGPT APIs. [10]

4.4 User Interface

Wildfire's user interface comprises three main components—a *task creation* page (Figure 4.3), a *task monitoring and expansion* page (Figure 4.4), and a *data analytics* page (Figure 4.6). The task creation page allows users to set up a task with various settings. The task monitoring and expansion page displays the status and progress of the tasks and allows users to control each task and configure its expansion. The data analytics page of each task allows users to explore the collected tweets using multiple filters and examine corresponding analytics results.

Task creation (Figure 4.3) Users have the option to create a task within one of three settings: search mode, tweet ID mode, or account mode. For the search mode, users can further toggle between simple (i.e., keywords) and advanced search

[10] https://platform.openai.com/docs/api-reference/chat

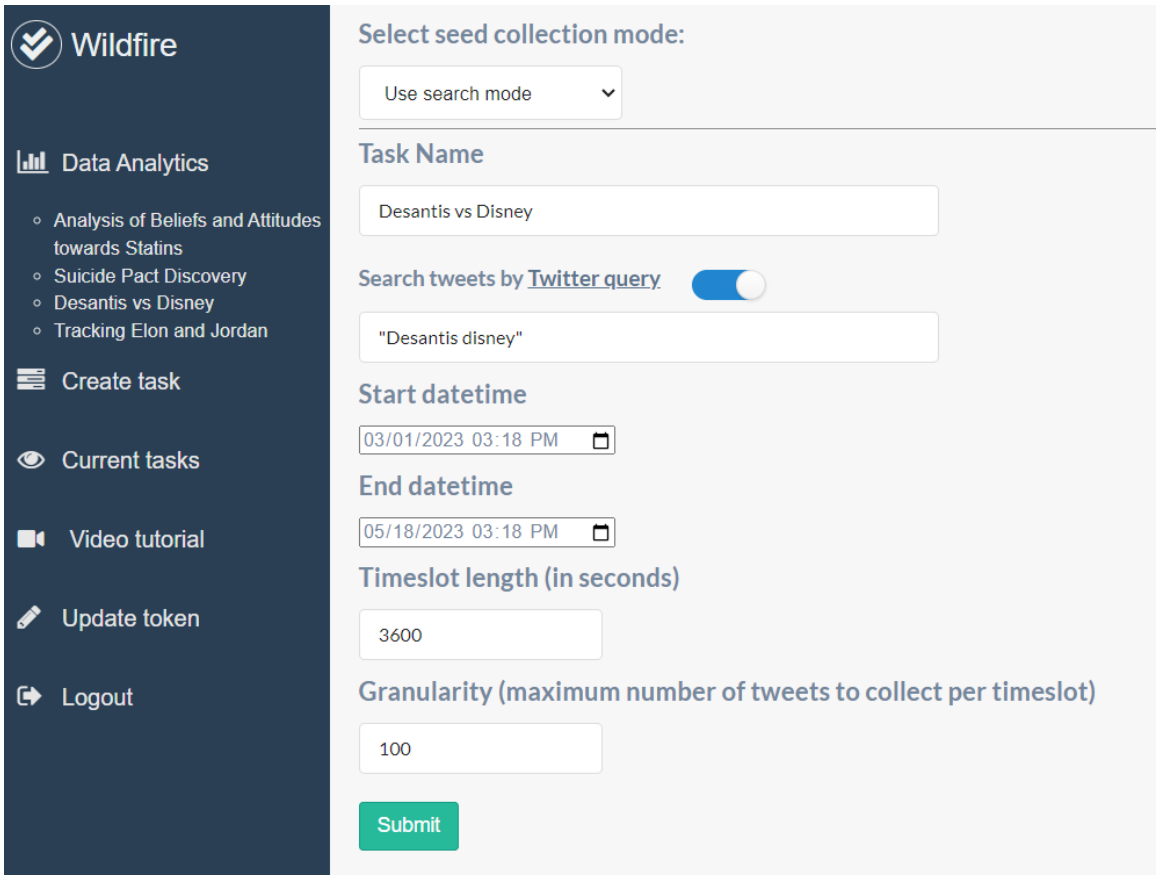


Figure 4.3: Task creation page

(i.e., query). The simple search allows users to enter multiple keywords separated by commas, in which a keyword may consist of multiple tokens. The advanced search allows users to enter a Twitter query. [9] If users already possess specific tweets or Twitter accounts of interest and intend to gather additional tweets using the ranking function, they can choose either tweet ID mode or account mode, where they must specify the tweet IDs or Twitter account handles. Users can also specify *Start datetime*, *End datetime*, *Timeslot length*, and *Granularity*, as discussed in Section 4.3.

Task monitoring and expansion (Figure 4.4) This page provides the overview of each task, displaying its configuration, status (such as active, completed, stopped, and error), and collection progress. For creating the task expansion, users

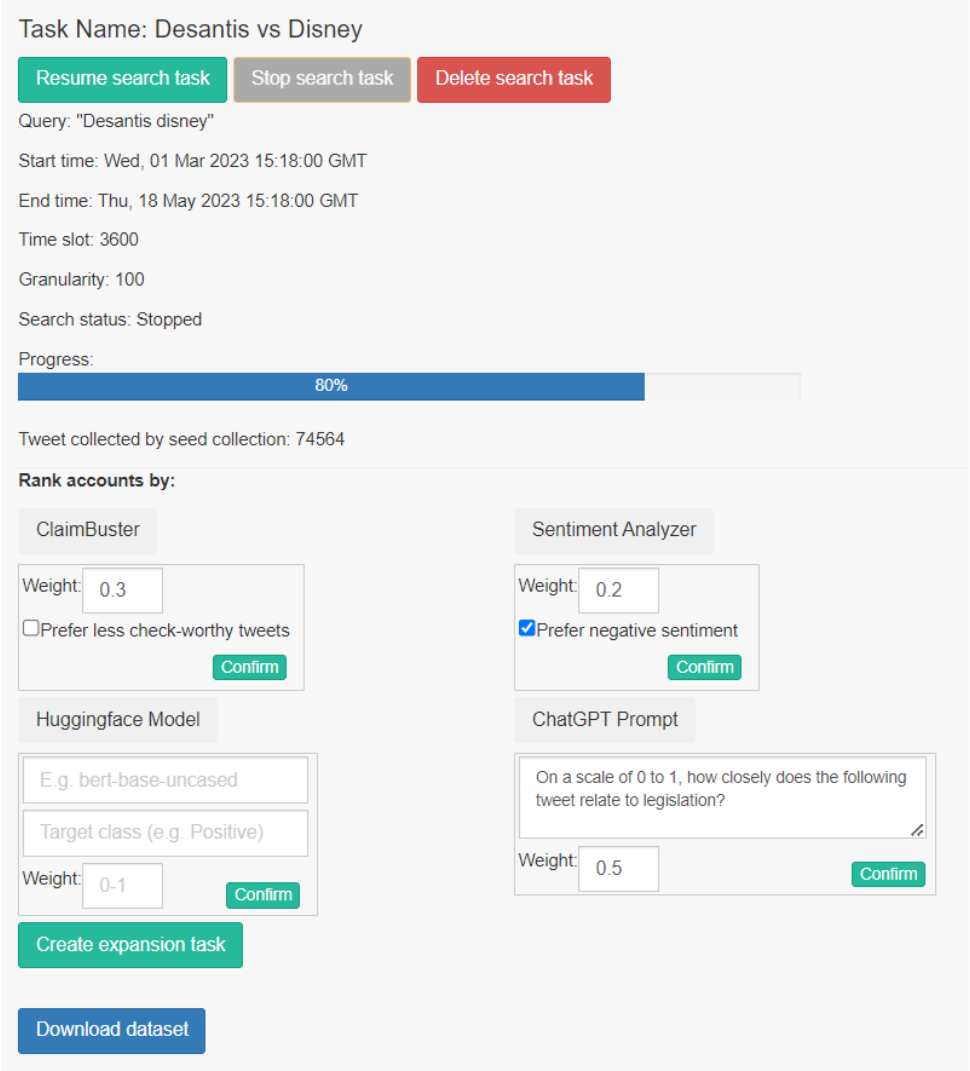


Figure 4.4: Task monitoring/expansion page

can choose the weight and preference associated with each ranking factor. Users are given options to start, stop, resume, or delete both seed and expansion collection tasks. The "Download" button at the bottom of each task leads to a page, as shown in Figure 4.5, where users can choose what to include in the downloaded dataset, e.g., whether to include the profile of each Twitter account.

Data analytics (Figure 4.6) To filter and analyze the collected data, Wildfire provides a search engine with various filtering options. Users can filter tweets based

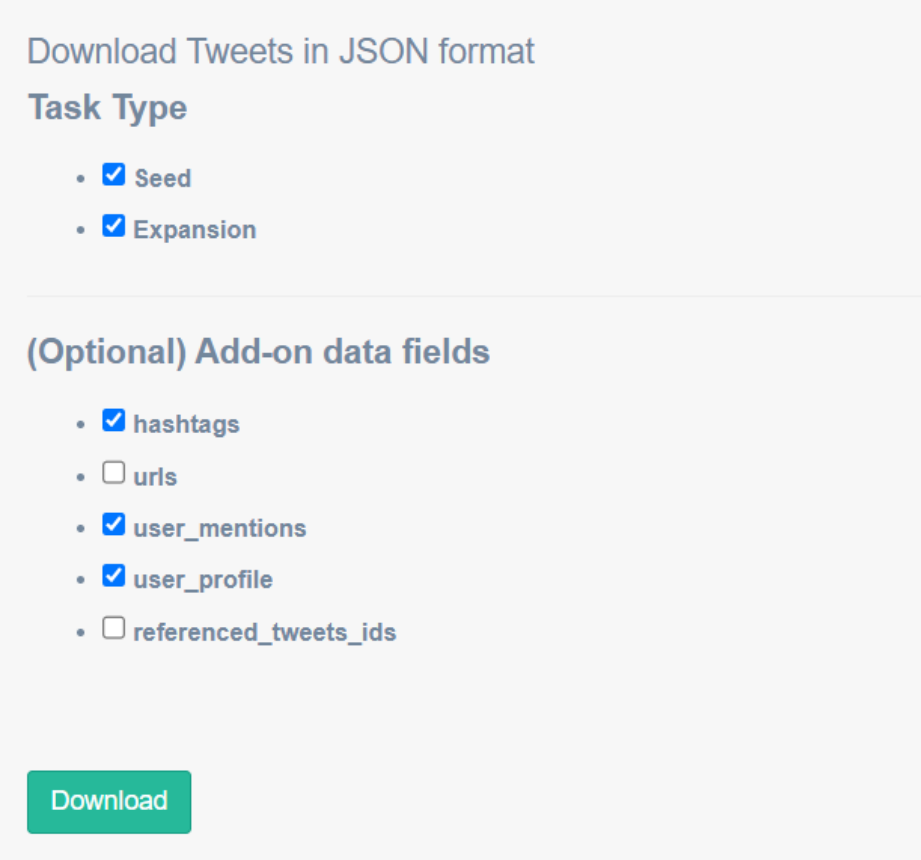


Figure 4.5: Dataset download page

on the task, keywords, hashtags, Twitter account handles, date range, and classification score range. Once the desired filters are applied, the corresponding result tweets are displayed and paginated. The analytic tools include topic generation [11] for generating thematic tags of tweets, entity recognition [12] for extracting entities from tweets, and the aforementioned two classifiers. Users have two ways to obtain analytic results. The first way is to click any tweet on the page, which opens a pop-up window (orange frame in Figure 4.6) showing the results of all four tools applied to the tweet. The output of the two classifiers is presented as scores between 0 and 1. The topic generation tool generates a list of topic words derived from the tweet text, although

[11] https://huggingface.co/fabiochiu/t5-base-tag-generation

[12] https://huggingface.co/autoevaluate/entity-extraction

these words do not necessarily appear in the tweet. The results of entity recognition are provided as pairs of entities and their corresponding entity types. Another way is to click one of the three result buttons: classifier results, topic generation results, and entity recognition results. The classifier results display the aggregated results derived from all the filtered tweets, including two classification score distributions presented as pie charts (the blue frame in Figure 4.6). The topic generation results show a bar chart displaying the top 10 most frequent tags representing the topics discussed among the tweets, and the entity recognition results reveal another bar chart showing the top 10 most frequent entities recognized from the tweets, as shown in Figure 4.7.

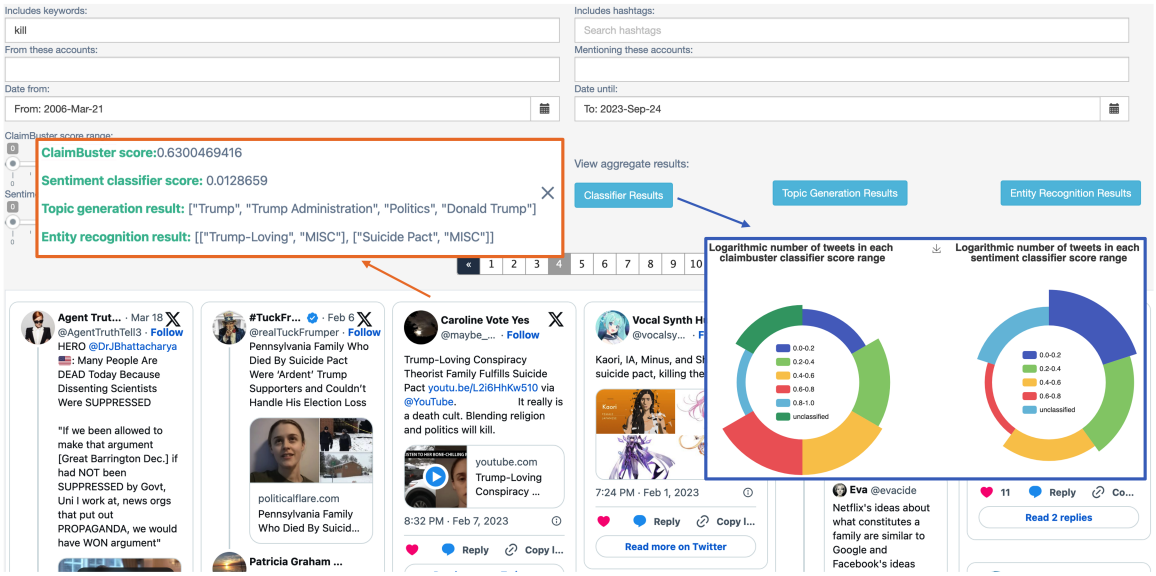
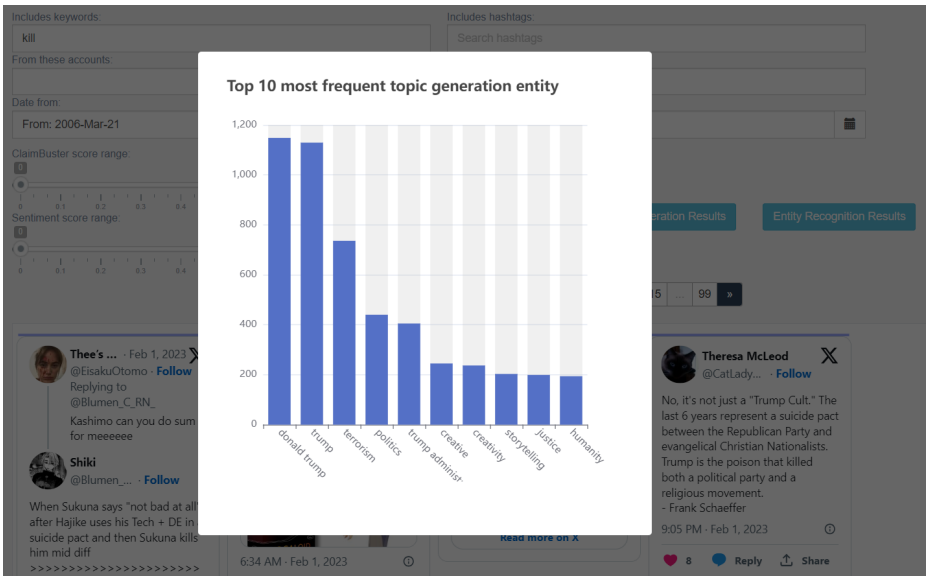


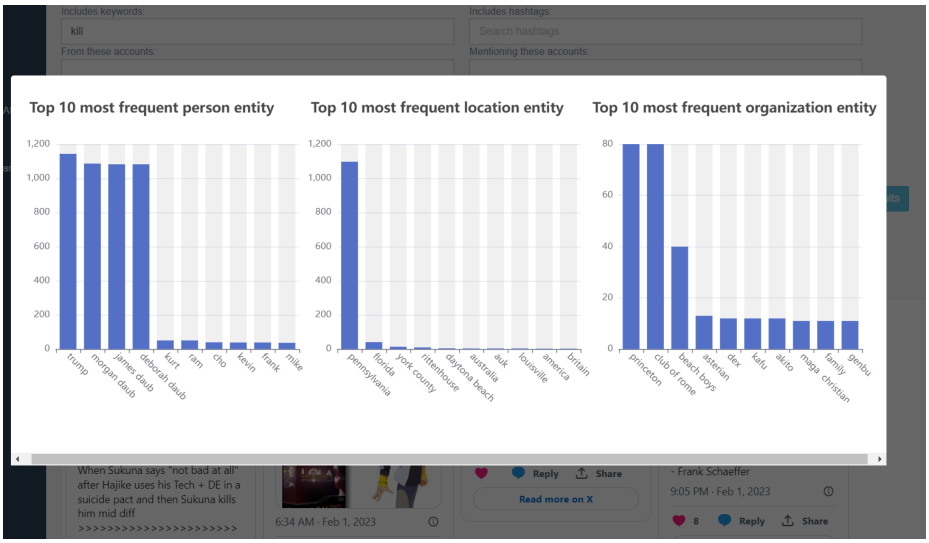
Figure 4.6: Data analytics page

4.5 Experiments

Many social sensing studies involve time-consuming data collection and analysis processes. It is high time to have an automated tool to aid such studies. Therefore,



(a) Topic generation results



(b) Entity recognition results

Figure 4.7: Example of aggregation results for topic generation and entity recognition

by presenting two use cases that replicate existing studies, we demonstrate the effectiveness of Wildfire in facilitating social sensing studies.

Use Case 1: Identifying political polarization on social media. To study political affiliations (e.g., left-wing or right-wing) on social media, researchers often use

keywords or hashtags to collect tweets and accounts in favor of or against certain political affiliations. For instance, Belcastro et al. [77] examined political polarization during the 2016 U.S. presidential election. The study used a dataset of millions of tweets collected by using keywords and hashtags that reflect support for two presidential candidates. Wildfire can help improve the data collection and analysis. We created a seed collection using keywords from the study and retrieved 4,347 tweets. By selecting the HuggingFace model politicalTweetBERT [13] in the ranking function, Wildfire found more tweets expressing support for Democrats. The expansion collection brought an additional 6,280 tweets. By comparing two samples of 100 tweets drawn from the seed collection and the expansion collection respectively, we found the high-scored tweets from the expansion were more relevant (67%) than those from the seed collection (48%), as illustrated by Table 4.1).

Method	**Total collected**	**Sample size**	**Accuracy**
Seed collection	4,347	100	48%
Expansion collection	6,280	100	67%

Table 4.1: Comparison of task relevance in seed and expansion collections

Use Case 2: How people make suicide pacts on Twitter. To investigate whether individuals utilize Twitter to search for like-minded persons with suicidal ideation and form suicide pacts, [78] collected tweets that contain "suicide pacts" in Korean from Oct. 16th to Nov. 30th, 2017. We used Wildfire to create a search collection with the same keywords in English and incorporated an expansion that applies Sentiment Analyzer and ChatGPT API with the prompt "On a scale of 0 to 1, how closely does the following tweet relate to discussions or mentions of suicide pacts?" in the ranking function. The expansion brought additional relevant tweets and accounts.

[13]https://api-inference.huggingface.co/models/m-newhauser/distilbert-political-tweets

The search collection retrieved 10,953 tweets and its expansion collected 24,112 tweets and 8 accounts from Feb 1st, 2023 to Mar 24th, 2023. Among the tweets from the expansion collection, 411 tweets are scored above 0.9 by ChatGPT. By manually checking 100-tweet samples from the 411 high-score tweets, we found that 79 tweets from 3 accounts are relevant to suicide. This indicates that the expansion equipped with ChatGPT effectively identified the target tweets.

CHAPTER 5

CASE STUDIES: SENSING THE SOCIETY WITH FACTUAL CLAIMS ON SOCIAL MEDIA

As discussed in Chapter 1, claim sensing offers a novel approach to understanding the complex dynamics between information dissemination and human interaction in online environments. In this chapter, we delve into the real-world practices of claim sensing, examining its applications. Through three case studies, we demonstrate the practical utility of claim sensing in domains such as misinformation debunking and public opinion analysis, highlighting the insights gained.

5.1 A Dashboard for Mitigating the COVID-19 Misinfodemic

Alongside the COVID-19 pandemic, there was a raging global misinfodemic [79, 80] just as deadly. As fear grew, false information related to the pandemic went viral on social media and threatened to affect an overwhelmed population. Such misinformation misled the public on how the virus was transmitted, how authorities and people were responding to the pandemic, as well as its symptoms, treatments, and so on. This onslaught exacerbated the vicious impact of the virus, as the misinformation drowned out credible information, interfered with measures to contain the outbreak, depleted resources needed by those at risk, and overloaded the health care system. Although health misinformation is not new [81], such a dangerous interplay between a pandemic and a misinfodemic was unprecedented. It calls for studying not only the outbreak but also its related misinformation; the fight on these two fronts must go hand-in-hand.

This study aims to understand the surveillance of, impact of, and effective interventions against the COVID-19 misinfodemic. To fulfill this, we built a dashboard that provides a map, a navigation panel, and timeline charts for looking up numbers of cases, deaths, and recoveries, similar to a number of COVID-19 tracking dashboards. [1][2][3] However, our dashboard also provides several features not found in other places. 1) It displays the most prevalent factual information among Twitter users in any user-selected U.S. geographic region. 2) The "factual information" comes from a catalog that we manually curated. It includes statements from authoritative organizations, verdicts, debunks, and explanations of (potentially false) factual claims from fact-checking websites, and FAQs from credible sources. The catalog's entries are further organized into a taxonomy. For simplicity, we refer to it as the *catalog and taxonomy of COVID-19 facts* or just *facts* in the ensuing discussion. 3) The dashboard displays COVID-19 related tweets from local authorities of user-selected geographic regions. 4) It embeds a chatbot built specifically for COVID-19 related questions. 5) It shows case-statistics from several popular sources which sometimes differ.

What is particularly worth noting about the underlying implementation of the dashboard is the adaptation of state-of-the-art textual semantic similarity and stance detection models. Tweets are first passed through a *claim-matching* model, which selects the tweets that semantically match the facts in our catalog. Then, the *stance detection* model determines whether the tweets agree with, disagree with, or merely discuss these facts. This enables us to pinpoint pieces of misinformation (i.e., tweets that disagree with known facts) and analyze their spread.

[1]https://www.covid19-trials.com

[2]https://coronavirus.jhu.edu/map.html

[3]https://www.cdc.gov/covid-data-tracker/index.html

A few studies analyzed and quantified the spread of COVID-19 misinformation on Twitter [82, 83, 84] and other social media platforms [85]. However, these studies conducted mostly manual inspection of small datasets, while our system automatically sifts through millions of tweets and matches tweets with our catalog of facts.

5.1.1 The Dashboard

Figure 5.1 shows the dashboard's user interface, with its components highlighted.

Geographic region selection panel (Component 1). A user can select a specific country, a U.S. state, or a U.S. county by using this panel or the interactive map (Component 2). Once a region is selected, the panel shows the counts of confirmed cases, deaths and recovered cases for the region in collapsed or expanded modes.

Interactive map (Component 2). On each country and each U.S. state, a red circle is displayed, with an area size proportional to its number of confirmed cases. When a state is selected, the circle is replaced with its counties' polygons in different shades of red, proportional to the counties' confirmed cases.

Timeline chart (Component 3). It plots the counts of the selected region over time and can be viewed in linear or logarithmic scale.

Panel of facts (Component 4). For the selected region, this panel displays facts from our catalog, and the distribution of people discussing, agreeing, or disagreeing with them on Twitter. A large number of people refuting these facts would indicate wide spread of misinformation. To avoid repeating misconceptions, the dashboard displays facts from authoritative sources only.

Government tweets (Component 5). It displays COVID-19 related tweets in the past seven days from officials of the user-selected geographic region. These tweets

are from a curated list of 3,744 Twitter handles that belong to governments, officials, and public health authorities at U.S. federal and state levels.

Chatbot (Component 6). This component embeds the *Jennifer Chatbot* built by the New Voices project of the National Academies of Sciences, Engineering and Medicine [86], which was built specifically for answering COVID-19 related questions.

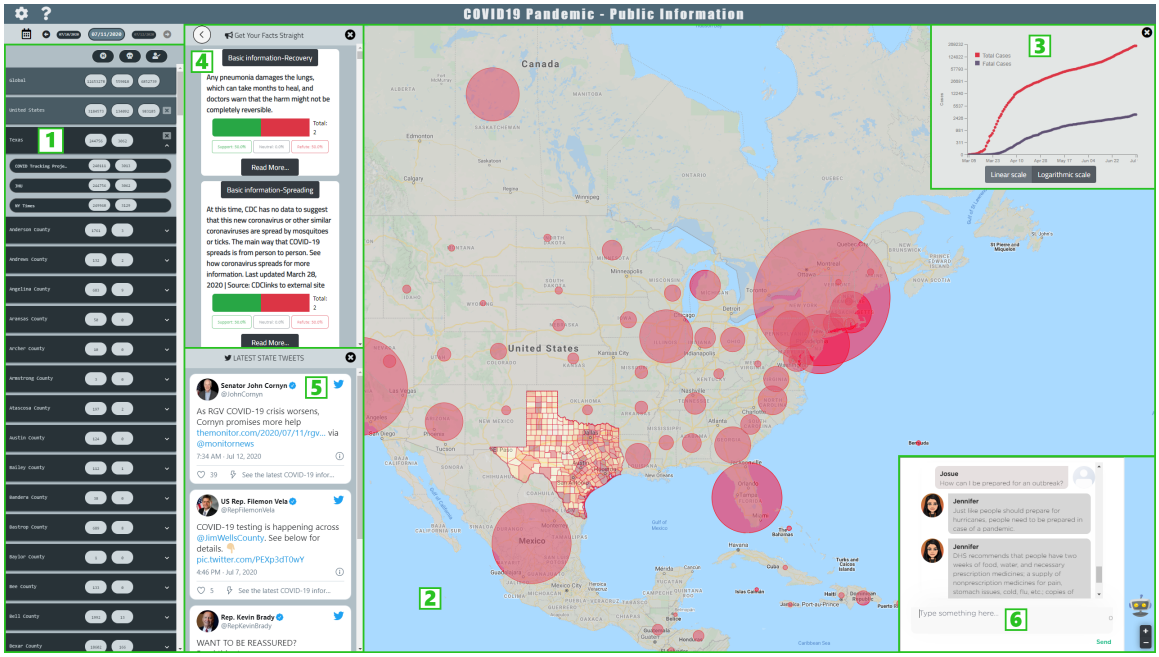


Figure 5.1: User interface of the dashboard for mitigating the COVID-19 misinfodemic

5.1.2 Datasets

The dashboard uses the following three datasets.

1) Counts of confirmed cases, deaths, and recoveries. We collected these counts daily from Johns Hopkins University, [4] the New York Times (NYT) [5] and the COVID

[4]https://github.com/CSSEGISandData/COVID-19

[5]https://github.com/nytimes/covid-19-data

Tracking Project. [6] These sources provide statistics at various geographic granularities (country, state, county).

2) Tweets. We use a collection of approximately 250 million COVID-19 related tweets from January 1st, 2020 to May 16th, 2020, obtained from [87] (version 10.0). We removed tweets and Twitter handles (and their tweets) that do not have location information, resulting in 34.6 million remaining tweets. We then randomly selected 10.4% of each month's tweets, leading to 3.6 million remaining tweets. We used the OpenStreetMap [7] API to map the locations of Twitter accounts from user-entered free text to U.S. county names. We used the ArcGIS API [8] to map the locations of tweets from longitude/latitude to counties.

3) A catalog and a taxonomy of COVID-19 related facts.

The manually curated catalog currently has 9,512 entries from 21 credible websites, including statements from authoritative organizations (e.g., WHO, CDC), verdicts, debunks, and explanations of factual claims (of which the truthfulness varies) from fact-checking websites (e.g., the IFCN CoronaVirusFacts Alliance Database, [9] PolitiFact), and FAQs both from credible sources (e.g., FDA, NYT) and a dataset curated by [88].

We organized the entries in this catalog into a taxonomy of categories, by integrating and consolidating the available categories from a number of source websites, placing entries from other websites into these categories or creating new categories, and organizing the categories into a hierarchical structure based on their inclusion relationship.

[6] https://covidtracking.com

[7] https://nominatim.openstreetmap.org

[8] https://developers.arcgis.com/python/guide/reverse-geocoding

[9] https://www.poynter.org/ifcn-covid-19-misinformation

Tweet	**Fact**	**Taxonomy Categories**	**Similarity**	**Stance**
Coronavirus cannot be passed by dogs or cats but they can test positive.	There has been no evidence that pets such as dogs or cats can spread the coronavirus.	Animals, Spreading	0.817	agree
More people die from the flu in the U.S. in 1 day than have died of the Coronavirus across the world ever.	Right now, it appears that COVID-19, the disease caused by the new coronavirus, causes more cases of severe disease and more deaths than the seasonal flu.	Cases	0.816	disagree

Table 5.1: Example results of matching tweets with facts and stance detection

5.1.3 Matching Tweets with Facts and Stance Detection

Given the catalog of COVID-19 related facts F and the tweets T, we first employ *claim-matching* to locate a set of tweets $\mathbf{t}^f \in T$ that discuss each fact $f \in F$. Next, we apply *stance detection* on pairs $\mathbf{p}^f = \{(t, f) \mid t \in \mathbf{t}^f\}$ to determine whether each t is agreeing with, disagreeing with, or neutrally discussing f. Finally, aggregate results are displayed on Component 4 of the dashboard to summarize the public's view on each fact. For both tasks, we employed Bidirectional Encoder Representations from Transformers (BERT) [89]. Table 5.1 shows some example results of claim matching and stance detection.

5.1.4 Misinformation Analysis

Figure 5.2 is the cumulative timeline for the top-6 countries with the most COVID-19 misinformation tweets in the dataset. "Misinformation tweets" refer to tweets that go against known facts as judged by our stance detection model.

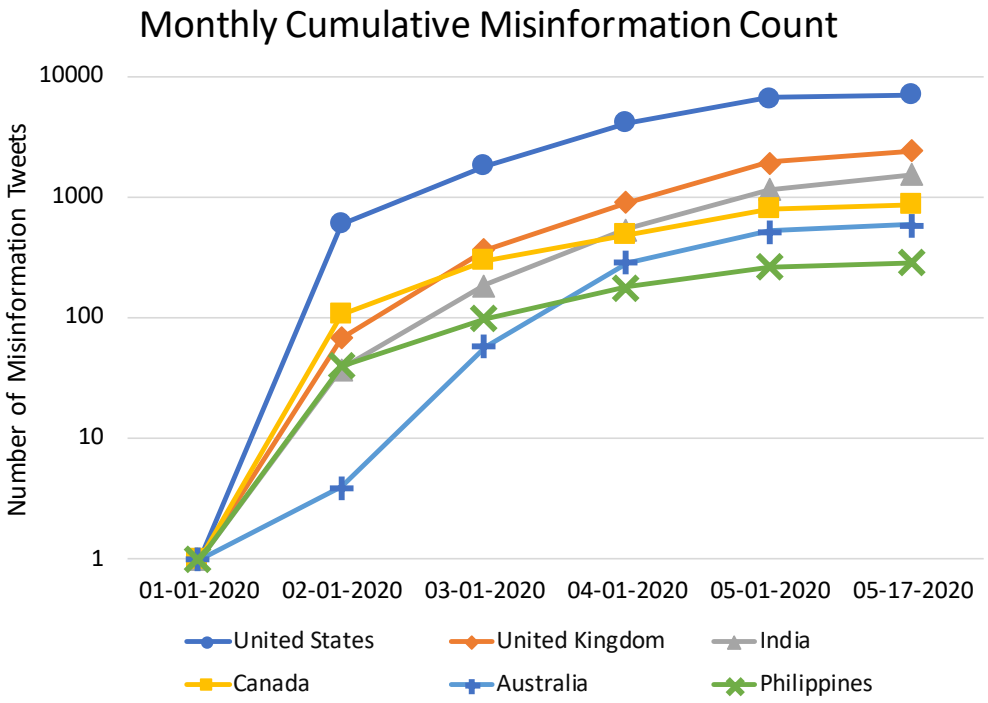


Figure 5.2: Six countries with the most misinformation tweets

We also conducted a study on the correlation between misinformation tweet counts and COVID-19 case counts. We looked at the percentage of cases relative to a country's population size, and the percentage of misinformation tweets relative to the total number of tweets from a country. The Pearson correlation coefficients between them are in Table 5.2. We find that the number of misinformation tweets most positively correlates with the number of confirmed cases. In contrast, its correlation with the number of recovered cases is weaker.

Finally, we manually categorized the misinformation tweets based on the taxonomy (Section 5.1.2). Table 5.3 lists the five most frequent categories of misinformation tweets. These five categories make up 49.9% of all misinformation tweets, with the other 50.1% being spread out over the other 33 categories.

Country	Confirm	Death	Recover
United States	0.763	0.738	0.712
United Kingdom	0.862	0.833	-
India	0.794	0.798	0.755
Canada	0.706	0.667	0.663
Australia	0.954	0.922	0.887
Philippines	0.720	0.696	0.618

Table 5.2: Correlation between the percentage of confirmed/deceased/recovered cases and the percentage of misinformation tweets. The number of recovered cases in U.K. after April 13th, 2020 is missing from the data source.

Category	Count	Percentage
Definition	2503	15.1
Spreading	2118	12.7
Other	1450	8.7
Testing	1301	7.8
Disease Alongside	936	5.6
Total	8308	49.9

Table 5.3: Most frequent categories of misinformation tweets

5.2 Granular Analysis of Social Media Users' Truthfulness Stances Toward Climate Change Factual Claims

Climate change is one of the most pressing global challenges of our time, profoundly impacting the environment, economy, and society. Amidst the urgency to address this global crisis, there is a large volume of discourse on climate change across social media platforms, reflecting growing public awareness and engagement. Understanding and analyzing discourse on climate change is crucial for informing public policy, media strategies, and societal awareness. Prior studies have explored various aspects of text analysis on climate change. [90] constructed a taxonomy of climate contrarian claims to analyze climate change myths and associated factual claims. Topic modeling performed on tweets by [91] showed that discussions of climate change span various topics. Stance detection [92, 93, 94] and sentiment analysis [95, 96] have

also been widely studied to understand people's beliefs and attitudes toward climate change.

In our study, we streamline a framework to sense people's truthfulness stances toward climate change on social media. In the framework, we first collect factual claims from five credible fact-checking websites using the keywords selected from the Environmental Protection Agency (EPA). Next, we gather corresponding social media posts using keywords extracted from the collected factual claims. We then leverage LLM with human-in-the-loop to automatically construct a climate change-related taxonomy. Finally, we fine-tune a truthfulness stance detection model to assess the truthfulness stances of social media posts toward their corresponding factual claims within the taxonomy. An overview of the framework is depicted in Figure 5.3.

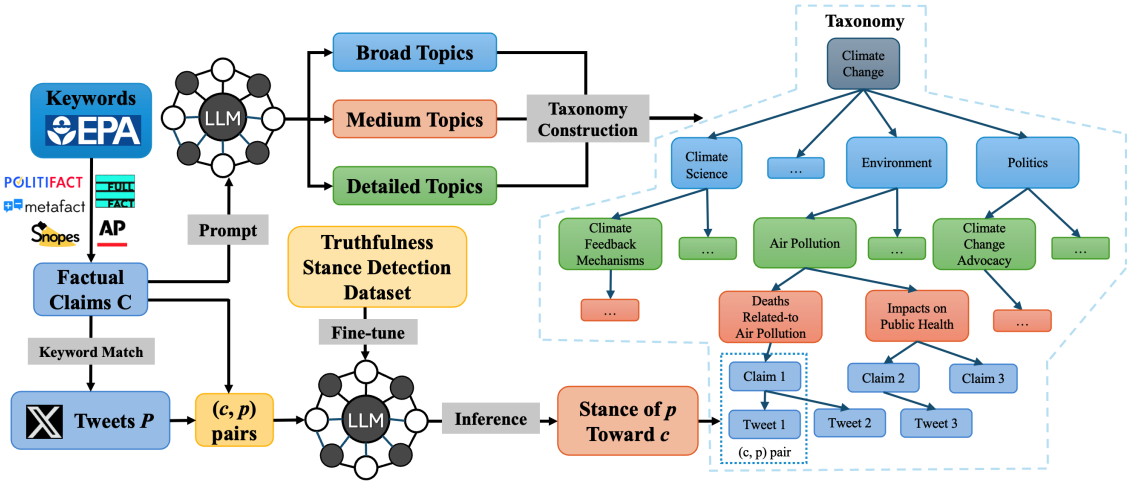


Figure 5.3: Overview of the framework for analyzing public judgments on climate change-related topics

5.2.1 Datasets

Factual Claim Collection To identify existing discourse related to climate change, we collect factual claims $\mathcal{C}$ from five fact-checking websites: *PolitiFact*, [10] *Snopes*, [11] *Full Fact*, [12] *Metafact*, [13] and *AP News*. [14] These websites are selected for their popularity and credibility in fact-checking. To collect $\mathcal{C}$, we curated a list of keywords related to climate change from the glossary of the Environmental Protection Agency (EPA) [15] to extract factual claims from the fact check websites. We consider a claim c to be climate change-related if any of the keywords appears in c itself, its fact-checking article's tags (i.e., the topics assigned to the article that categorize its content), or the articles' content. We also collected the verdicts of $\mathcal{C}$ (e.g., "Mostly-true," "False") determined by the fact-checking websites. After removing duplicates, we obtained $1,409$ unique climate change-related factual claims spanning from November 2007 to May 2024.

Tweet Collection After identifying existing climate change-related factual claims $\mathcal{C}$, we collected corresponding tweets $\mathcal{P}$ discussing those claims to construct (c, p) pairs. This allows us to assess people's judgments of different claims, i.e., whether the tweet c believes the factual claim p is true or false. To construct (c, p) pairs, we used the tokens extracted from c to collect relevant tweets that discuss each c from X. In this way, we collected a total of $13,050$ tweets for 729 out of $1,409$ claims. Among these 729 claims, 294 claims had more than 10 tweets.

[10] politifact.com
[11] snopes.com
[12] fullfact.org
[13] metafact.io
[14] apnews.com
[15] EPA Keyword List

5.2.2 Methodologies

Taxonomy Construction A taxonomy serves as a hierarchical classification structure, organizing topics from broader to more fine-grained levels of granularity. In this framework, we aim to generate a three-level taxonomy from factual claims $\mathcal{C}$ related to climate change. To enhance the efficiency of the taxonomy-building process, we prompt LLM, specifically Zephyr [97], to generate a set of broad topic, medium topic, and detailed topic, denoted as $\{t^b, t^m, t^d\}$, for each factual claim $c \in \mathcal{C}$. We adopt human-in-the-loop to refine the prompt based on the generated topics, enabling multi-round topic generation for optimal results. More specifically, after the LLM produces $\{\hat{t_j^b}, \hat{t_j^m}, \hat{t_j^d}\}$ for all $c_j \in \mathcal{C}$, humans scrutinize broad topics that appear frequently (i.e., more than 40 times), identify new topic sets that both include those frequent broad topics and accurately represent their associated claims, and construct the taxonomy according to the topic hierarchy. The new topic sets and associated claims are used as new learning examples for the next round of topic generation and taxonomy construction, continuing until no new frequent broad topics are generated.

Truthfulness Stance Detection The task of truthfulness stance detection [98] involves determining the stance of a social media post p toward a factual claim c. The stance can be classified as either believing c is true (*Positive* ($\oplus$)), believing c is false (*Negative* ($\ominus$)), or expressing a neutral stance or no stance toward c (*Neutral/No stance* ($\odot$)). We apply supervised fine-tuning to an LLM to build a classifier, leveraging Zephyr [97] as the underlying backbone LLM, and we use an in-house annotated dataset that contains claim-tweet pairs (c, p) and stance labels as the ground truth.

Claim	Tweet	Stance	Broad Topic	Medium Topic	Detailed Topic
Air pollution linked to greater risk of dementia.	People over 50 in areas with the highest levels of nitrogen oxide in the air showed a 40% greater risk of developing dementia than those with the least NOx #airpollution.	⊕	Health	Air Pollution	Impacts on Brain Health
Sen. Lindsey Graham supports the Green New Deal.	Facebook removed an ad by Adriel Hampton showing Sen. Lindsey Graham backing the Green New Deal.	⊙	Politics	Climate Change Advocacy	Politicians' Stance
The Earth is warming because of the sun's changing distance from the Earth, not because of carbon emissions.	Enough with your pseudo-scientific. Actual science has proven the relationship to human carbon emissions and not cycles of sun /earth distance.	⊖	Climate Science	Climate Feedback Mechanisms	Misconceptions

Table 5.4: Examples of truthfulness stance detection and their corresponding topics in the taxonomy

Broad Topic	Truth-⊕	Truth-⊖	Misi-⊕	Misi-⊖	Accuracy	Macro F1
Climate Science	81.7% (524)	18.3% (117)	72.5% (377)	27.5% (143)	0.575	0.524
Economy	70.5% (146)	29.5% (61)	72.5% (351)	27.5% (133)	0.404	0.404
Energy	82.2% (264)	17.8% (57)	74.7% (124)	25.3% (42)	**0.628**	**0.530**
Environment	77.5% (533)	22.5% (155)	74.4% (1040)	25.6% (357)	0.427	0.423
Government Policies	83.2% (183)	16.8% (37)	**69.5%** (205)	**30.5%** (90)	0.530	0.514
Health	**88.7%** (180)	**11.3**% (23)	**77.9%** (169)	**22.1%** (48)	0.543	0.493
Politics	**69%** (363)	**31%** (163)	75.7% (1635)	24.3% (525)	**0.331**	**0.329**
Technology	74.8% (86)	25.2% (29)	69.8% (120)	30.2% (52)	0.481	0.473

Table 5.5: Stance distribution towards **Truth** and **Misi**information across broad topics. Truth-⊕ and Truth-⊖ denote positive and negative stances towards **Truth**, respectively. Misi-⊕ and Misi-⊖ denote positive and negative stances towards **Misi**information, respectively. Note that the topic "*Others*" is not considered in this analysis.

5.2.3 Results

Table 5.4 presents examples of the topics generated in the taxonomy and the stances identified by the truthfulness stance detection model for each (c, p) pair. Each pair is associated with the tweet's truthfulness stance toward the claim, while the topics are used to describe the claim. To explore whether social media users can discern true and false claims on various climate change-related topics, we calculated the distribution of positive and negative stances in tweets toward claims with verified

verdicts of either true (Truth) or false (Misinformation), as presented in Table 5.5. We also calculated accuracy to examine how the stances align with the claims' veracity. In addition to accuracy, the macro F1 score was chosen due to the imbalance in the claims' verdicts. We excluded claims from "*Others*" for their small sample size, as well as claims with "Uncertain" verdict and tweets classified as $\odot$, as they provide less meaningful insights.

The high percentage of both Truth-$\oplus$ and Misi-$\oplus$ suggests that people tend to believe claims are true regardless of their actual truthfulness. Furthermore, people are more likely to believe claims related to "*Health*," given it has the highest Truth-$\oplus$ (88.7%) and Misi-$\oplus$ (77.9%). The variation in accuracy and macro F1 scores across different topics indicates that people's judgments vary significantly depending on the topics. The low accuracy and macro F1 scores reveal that social media users' judgments of factual claims are not very accurate in the broad topics of "*Politics*" (0.331, 0.329), "*Economy*" (0.404, 0.404), and "*Environment*" (0.427, 0.423) (Table 5.5). The highest macro F1 score is 0.53 for "*Energy*," while most topics' macro F1 score is below 0.5. This suggests that social media users struggle to distinguish between true and false claims. This finding is consistent with the results reported by [99] in social science, which suggest that social media users have difficulty detecting fake news and that most users would make more accurate judgments by simply flipping a coin.

5.3 Evaluating the Impact of Check-Worthiness on Retweet Prediction Models

In contemporary society, online social networks are integral to our everyday routines. These digital platforms link individuals based on social connections, professional ties, or shared interests across diverse topics. Users can publish updates about their hobbies, views, and experiences, while engaging with others through comments, likes, and shares. By understanding what information users are inclined to share or

spread, one can develop more effective marketing strategies, optimize advertisement channels, and enhance social network designs for better information dissemination and user engagement. Hence, it is crucial to study the patterns of information diffusion and predict when it will occur. This study focuses on Twitter, where retweeting is the primary method of information diffusion. A retweet involves reposting a tweet originally posted by another user. By predicting if a tweet will be retweeted, we can determine if a user might spread the information within the tweet. Since users play a central role in the process of information diffusion, it is essential to comprehend their preferences for spreading information (including what, when, and whether to spread) by analyzing their past behaviors. Understanding these preferences aids in predicting retweets, determining if a tweet might go viral, or finding the best way to disseminate information to reach a broader audience.

Various explicit and implicit content-based features, along with features related to social connections, have been explored for retweet prediction [100, 101, 102, 103, 104, 105, 106, 107]. Content-based features pertain to the information in the tweet's text, while social features relate to the relationships between users (the author and the retweeter). Explicit content features, such as hashtags, URLs, emoticons, mentions, locations, etc. are directly accessible. Implicit features such as sentiments, political spectrum, age brackets, interests, etc. require tools or algorithms for extraction such as TFIDF-weighted terms [108], LDA topics [109], topic novelty [110], sentiment analysis [111], and emotional divergence [112]. Social features consider whether one user follows another, if the follow relationship is mutual, and the frequency and nature of their interactions (replies or retweets). These features provide insights into the potential influence of the tweet's author on the retweeter's decision and the extent of that influence.

Although many researchers have studied various features related to retweet prediction, none of them has examined check-worthiness. Since Chapter 3 demonstrated that people have the behavioral tendency to post, retweet, and like content with similar check-worthiness levels, it is worth examining whether check-worthiness can be a good feature for retweet prediction models. Hence, in this study, we aim to explore the effect of check-worthiness on users' retweet activities. More specifically, we study whether including check-worthiness as one of the features can improve the performance for the following modeling task:

Given a Twitter account X and a target tweet T, predict whether account X will retweet tweet T.

5.3.1 Methodology

In this study, we employ machine learning (classification) techniques to build our prediction models and compare their accuracies. Retweet prediction is inherently a prediction problem, making machine learning methods a natural choice for addressing it. While there are numerous machine learning models available for comparison, this study focuses exclusively on the BERT model [89] and its variants. Specifically, we utilize BERT models fine-tuned on Twitter data, such as Bertweet [113], Twhin-bert [114], and Twitter-XML-RoBERTa [75]. We selected BERT for its benchmarking status in various NLP classification tasks and opted for these specific variants because their fine-tuning on Twitter data allows for a more accurate understanding of tweet content.

Model Architecture Our model architecture is shown in Figure 5.4. The proposed model for retweet prediction is able to leverage the input data including:

- **Target Tweet**: The tweet to predict whether the user will retweet or not.

- **User Profile**: The user's profile description written on their homepage.
- **User Historical Retweets**: Tweets that were retweeted by the user.
- **Check-worthiness**: The difference between the user's check-worthiness (i.e., the individual check-worthiness mentioned in Section 3.6) and the target tweet's check-worthiness.

Each input type is passed through an embedding model (e.g., BERT) to generate embeddings or vector representations. Historical retweets are individually embedded and then aggregated into a single representation. This aggregated embedding, along with embeddings for the user profile, target tweet, and check-worthiness difference, are combined into a unified feature vector. A dropout layer followed by a fully connected (FC) layer is applied to the historical retweets' embeddings to mitigate overfitting and transform the aggregated representations. The combined feature vector is then passed through another FC layer to produce the final feature vector, which is subsequently fed into a classifier to generate the retweet prediction. This architecture effectively integrates multiple sources of information to enhance the accuracy of retweet predictions.

Dataset We conducted experiments on the POL dataset from Section 3.4. We selected the POL dataset because most users in it are politically oriented and have more interactions with each other, providing a wealth of following relationships and interactions for analysis. While the original POL dataset includes users and their tweets, it lacks information on their following relationships. Thus, we collected additional data on these relationships. For training and testing the models, a ground truth dataset is necessary. The POL dataset provides up to 3,200 recent tweets in each user's timeline, including retweets, which serve as our positive samples. Negative samples are harder to identify because not all tweets that a user did not retweet can

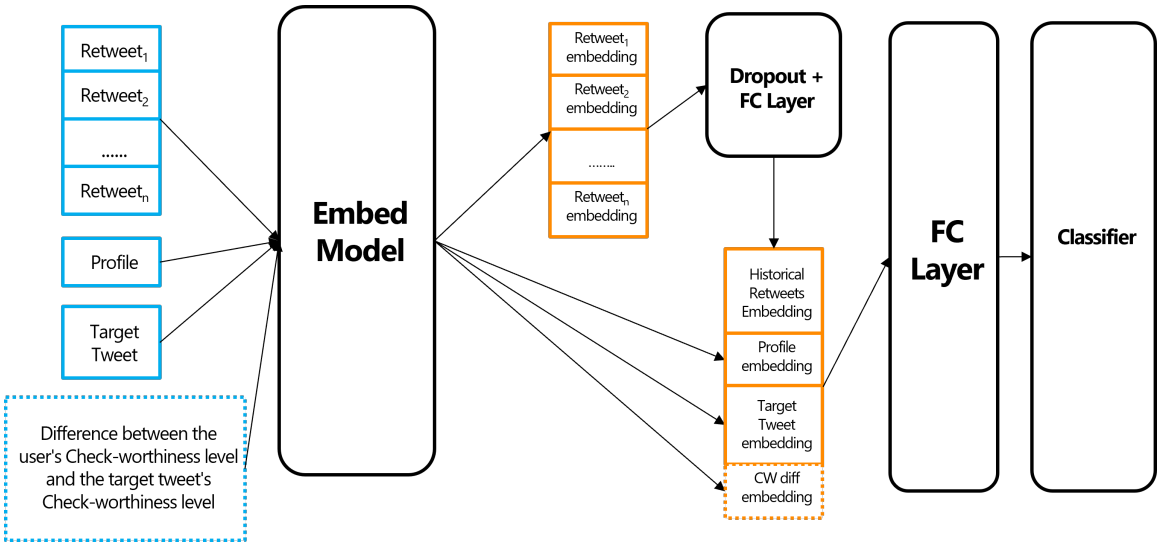


Figure 5.4: Architecture of the proposed retweet prediction model

be considered negative samples—they might simply have not been seen by the user. For a tweet to qualify as a negative sample, it must have been seen by the user who chose not to retweet it. However, it is impossible to determine if a user has seen a specific tweet. To address this, we leverage the following relationships. We assume that if a user has retweeted one of their followees more than twice, they are likely paying attention to that followee and have seen all of their tweets. Consequently, we derive negative samples by extracting all the tweets from the followee that the user did not retweet. Figure 5.5 provides an example. All tweets retweeted by user U are considered positive samples, regardless of the original authors. However, only if user U has retweeted its followee F's tweets more than twice (e.g., $Tweet1$ and $Tweet4$), the remaining tweets from F that U has not retweeted (e.g., $Tweet2$ and $Tweet3$) can be treated as negative samples. Using this method, we generated data for 500 users, each with 50 to 200 positive samples and 30 to 100 negative samples, totaling 125,911 samples. For each user, the earliest 40 positive samples (i.e., its chronologically earliest retweets) were reserved as the User Historical Retweets feature, while

the remaining positive samples were used as ground truth for training and testing. All negative samples were also treated as ground truth. Of the ground truth data, 90% were used for training, and the remaining 10% were used for testing.

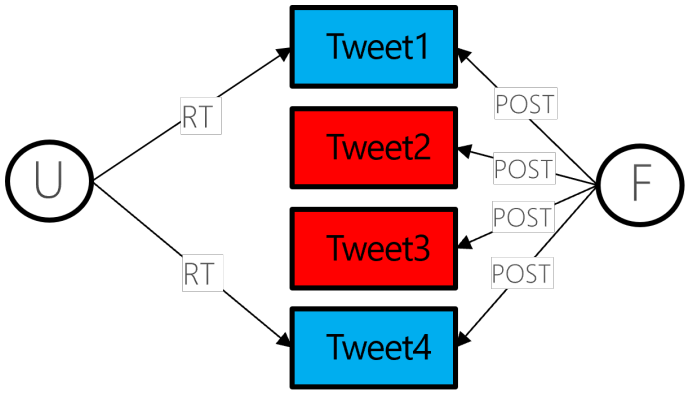


Figure 5.5: Example of producing positive (blue) and negative (red) data samples

5.3.2 Experiments

We fine-tuned 4 aforementioned models (i.e., BERT, BERTweet, Twhin-BERT, and Twitter-XML-RoBERTa) on the dataset mentioned above for $8,394$ steps with a batch size of 6 using the Adam optimizer [115] and an initial learning rate of 2×10^{-5}. We employed a linearly decreasing learning rate schedule without warm-up. Our implementation was based on PLMs from Hugging Face [116]. For each model, we trained and tested using 4 different sets of features:

- **Profile**: User Profile only
- **Profile+CW**: User Profile and Check-worthiness
- **Profile+Hist.RTs**: User Profile and Historical Retweets
- **Profile+Hist.RTs+CW**: User Profile, Historical Retweets, and Check-worthiness

The results are presented in Table 5.6. With the same feature settings, it can be observed that the inclusion of the check-worthiness feature slightly improves the

model's performance across all models. However, this improvement is minimal and may be considered negligible in practical terms. One possible explanation for this outcome is that the language models employed are already effective at capturing the nuances of check-worthiness within the tweet text itself. As a result, the explicit addition of the check-worthiness feature does not provide the substantial boost in predictive power that we initially anticipated.

	Profile	**Profile+CW**	**Profile+Hist.RTs**	**Profile+Hist.RTs+CW**
BERT	64.6%	66.1%	70%	71%
BERTweet	76.4%	78.1%	80%	80.3%
Twhin-BERT	78.1%	79.3%	83.4%	83.8%
Twitter-XML-RoBERTa	67.3%	68.1%	72.2%	72.5%

Table 5.6: Performance of fine-tuned language models

CHAPTER 6

CONCLUSION

In this study, we introduce and explore the concept of claim sensing, delving into the connections between factual claims and human behaviors on social media. Our work offers a fresh perspective on observing societal dynamics and provides valuable insights into how factual claims shape people's preferences, interactions, and ideologies on these platforms. We present experiments that investigate behavioral tendencies toward factual claims, along with the methods and tools that facilitate social sensing studies. Furthermore, we highlight real-world examples of sensing public opinion and investigating information dissemination behaviors on social media through the lens of factual claims.

The work of Chapter 3 explores correlations between human behaviors and check-worthiness and identifies the existence of the difference between individuals' behavioral tendencies toward factual claims. Through a meticulously designed set of experiments, the research has established a correlation between these tendencies and users' posting, sharing, and liking behaviors, indicating a consistent tendency to engage with content of similar check-worthiness. Furthermore, the findings emphasize the heightened efficacy of two-way following relationships in reflecting shared preferences towards factual claims. The concept of check-worthiness emerges as a potent tool for understanding human behaviors within the realm of social media. Our results not only provide valuable insights into the impact and adaptability of check-worthiness but also lay the groundwork for future investigations to delve deeper into

the various dimensions of its influence on social media behaviors and its potential applications across diverse domains.

The work of Chapter 4 presents Wildfire, an innovative social sensing platform designed for laypersons, which supports users in conducting social sensing tasks using Twitter data without programming and data analytics skills. Wildfire employs a heuristic graph exploration method to selectively expand the collected tweet-account graph in order to further retrieve more task-relevant tweets and accounts. This approach allows for the collection of data to support complex social sensing tasks that cannot be met with a simple keyword search. In addition, Wildfire provides a range of analytic tools, such as text classification, topic generation, and entity recognition, which can be crucial for tasks such as trend analysis. The platform also provides a web-based user interface for creating and monitoring tasks, exploring collected data, and performing analytics. Two case studies were performed to validate the effectiveness of Wildfire.

The work of Chapter 5 presents real-world cases of sensing public opinions, debunking misinformation, and predicting retweeting behaviors through the perspective of factual claims. We found that people tend to accept claims as true, regardless of their accuracy. This issue is particularly evident in discussions on politics, economy, and environment. We discovered the flow of misinformation and public stances toward those misinformation and identified frequent categories of misinformation tweets during the Covid pandemic period. We evaluated the impact of check-worthiness feature on retweet prediction models. Those practices and findings show that sensing social media using factual claims can bring unique and useful insights into our society.

Claim sensing is a broad and multifaceted topic, and the studies presented here demonstrate its practices and value. However, these represent only a small part of the broader field, leaving much to be explored in the future. For instance, building on the

concept of check-worthiness, future research could develop more refined metrics and tools to assess the credibility and relevance of claims. This might involve integrating machine learning models that automatically evaluate both check-worthiness and validity, making it a more dynamic and adaptable measure for understanding human behavior and information flow on social media. Additionally, conducting longitudinal studies to track changes in user behavior and misinformation patterns over time could provide valuable insights into how social media dynamics evolve in response to significant events, such as elections or pandemics. This could aid in predicting future trends and developing strategies to mitigate the spread of misinformation. Moreover, expanding the application of claim sensing to fields such as education, healthcare, and marketing could be a promising direction. Understanding how factual claims influence decision-making processes in these areas might offer new opportunities for applying claim sensing across diverse domains. In terms of data collection and analytics, future work could explore the integration of Wildfire with other social media platforms beyond Twitter, such as Facebook, Instagram, or emerging platforms such as Threads. This would enable a more comprehensive analysis of social sensing tasks and lead to more robust, cross-platform insights.

REFERENCES

[1] B. I. Page and R. Y. Shapiro, "Effects of public opinion on policy," *American political science review*, vol. 77, no. 1, pp. 175–190, 1983.

[2] J. Pacheco, "The social contagion model: Exploring the role of public opinion on the diffusion of antismoking legislation across the american states," *The Journal of Politics*, vol. 74, no. 1, pp. 187–202, 2012.

[3] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nature communications*, vol. 10, no. 1, p. 7, 2019.

[4] F. P. Barclay, P. Chinnasamy, and P. Pichandy, "Political opinion expressed in social media and election outcomes-us presidential elections 2012," *GSTF International Journal on Media & Communications (JMC)*, vol. 1, no. 2, pp. 15–22, 2014.

[5] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 us presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[6] J. Brummette, M. DiStaso, M. Vafeiadis, and M. Messner, "Read all about it: The politicization of "fake news" on twitter," *Journalism & Mass Communication Quarterly*, vol. 95, no. 2, pp. 497–517, 2018.

[7] N. Hassan, M. Tremayne, F. Arslan, and C. Li, "Comparing automated factual claim detection against judgments of journalism organizations," in *Computation+ journalism symposium*, 2016, pp. 1–5.

[8] S. Majithia, F. Arslan, S. Lubal, D. Jimenez, P. Arora, J. Caraballo, and C. Li, "Claimportal: Integrated monitoring, searching, checking, and analytics of fac-

tual claims on twitter," in *Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations*, 2019, pp. 153–158.

[9] M. Samadi, P. Talukdar, M. Veloso, and M. Blum, "Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[10] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.

[11] N. Hassan, C. Li, and M. Tremayne, "Detecting check-worthy factual claims in presidential debates," in *Proceedings of the 24th acm international on conference on information and knowledge management*, 2015, pp. 1835–1838.

[12] B. Adair, C. Li, J. Yang, and C. Yu, "Progress toward "the holy grail": The continued quest to automate fact-checking," in *Computation+ Journalism Symposium,(September)*, 2017.

[13] Y. Godler and Z. Reich, "Journalistic evidence: Cross-verification as a constituent of mediated knowledge," *Journalism*, vol. 18, no. 5, pp. 558–574, 2017.

[14] J. M. W. Lewis, A. Williams, R. A. Franklin, J. Thomas, and N. A. Mosdell, "The quality and independence of british journalism," 2008.

[15] T. Flew, C. Spurgeon, A. Daniel, and A. Swift, "The promise of computational journalism," *Journalism practice*, vol. 6, no. 2, pp. 157–171, 2012.

[16] D. Graves, "Understanding the promise and limits of automated fact-checking," *Reuters Institute for the Study of Journalism*, 2018.

[17] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *Acm Computing Surveys (Csur)*, vol. 51, no. 2, pp. 1–36, 2018.

[18] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, no. 1, p. 82, 2020.

[19] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.

[20] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "A survey on stance detection for mis-and disinformation identification," *arXiv preprint arXiv:2103.00242*, 2021.

[21] N. Kotonya and F. Toni, "Explainable automated fact-checking: A survey," *arXiv preprint arXiv:2011.03870*, 2020.

[22] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, and G. D. S. Martino, "Automated fact-checking for assisting human fact-checkers," *arXiv preprint arXiv:2103.07769*, 2021.

[23] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," *arXiv preprint arXiv:1811.00770*, 2018.

[24] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[25] G. D. S. Martino, S. Cresci, A. Barrón-Cedeño, S. Yu, R. Di Pietro, and P. Nakov, "A survey on computational propaganda detection," *arXiv preprint arXiv:2007.08024*, 2020.

[26] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[27] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[28] F. Cardoso Durier da Silva, R. Vieira, and A. C. Garcia, "Can machines learn to detect fake news? a survey focused on social media," 2019.

[29] T. S. Rich, I. Milden, and M. T. Wagner, "Research note: Does the public support fact-checking social media? it depends who and how you ask," *The Harvard Kennedy School Misinformation Review*, 2020.

[30] P. B. Brandtzaeg, A. Følstad, and M. Á. Chaparro Domínguez, "How journalists and social media users perceive online fact-checking and verification services," *Journalism practice*, vol. 12, no. 9, pp. 1109–1129, 2018.

[31] J. Zhang, J. D. Featherstone, C. Calabrese, and M. Wojcieszak, "Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines," *Preventive Medicine*, vol. 145, p. 106408, 2021.

[32] S. Jiang and C. Wilson, "Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–23, 2018.

[33] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan *et al.*, "Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media," *Political behavior*, vol. 42, pp. 1073–1095, 2020.

[34] A. Carson, T. B. Gravelle, J. B. Phillips, J. Meese, and L. Ruppanner, "Do brands matter? understanding public trust in third-party factcheckers of misinformation and disinformation on facebook," *International Journal of Communication*, vol. 17, p. 25, 2023.

[35] M. S. Islam, T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar *et al.*, "Covid-19–related infodemic and its impact on public health: A global social media analysis," *The American journal of tropical medicine and hygiene*, vol. 103, no. 4, p. 1621, 2020.

[36] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

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

[38] J. M. Carey, A. M. Guess, P. J. Loewen, E. Merkley, B. Nyhan, J. B. Phillips, and J. Reifler, "The ephemeral effects of fact-checks on covid-19 misperceptions in the united states, great britain and canada," *Nature Human Behaviour*, vol. 6, no. 2, pp. 236–243, 2022.

[39] A. L. Wintersieck, "Debating the truth: The impact of fact-checking during electoral debates," *American politics research*, vol. 45, no. 2, pp. 304–331, 2017.

[40] S. J. Newell, R. E. Goldsmith, and E. J. Banzhaf, "The effect of misleading environmental claims on consumer perceptions of advertisements," *Journal of Marketing Theory and Practice*, vol. 6, no. 2, pp. 48–60, 1998.

[41] T. G. Van der Meer and Y. Jin, "Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source," *Health communication*, vol. 35, no. 5, pp. 560–575, 2020.

[42] N. Hassan, G. Zhang, F. Arslan, J. Caraballo, D. Jimenez, S. Gawsane, S. Hasan, M. Joseph, A. Kulkarni, A. K. Nayak *et al.*, "Claimbuster: The first-

ever end-to-end fact-checking system," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1945–1948, 2017.

[43] C. Hansen, C. Hansen, S. Alstrup, J. Grue Simonsen, and C. Lioma, "Neural check-worthiness ranking with weak supervision: Finding sentences for fact-checking," in *Companion proceedings of the 2019 world wide web conference*, 2019, pp. 994–1000.

[44] S. Vasileva, P. Atanasova, L. Màrquez, A. Barrón-Cedeño, and P. Nakov, "It takes nine to smell a rat: Neural multi-task learning for check-worthiness prediction," *arXiv preprint arXiv:1908.07912*, 2019.

[45] D. Wright and I. Augenstein, "Claim check-worthiness detection as positive unlabelled learning," *arXiv preprint arXiv:2003.02736*, 2020.

[46] C. Lespagnol, J. Mothe, and M. Z. Ullah, "Information nutritional label and word embedding to estimate information check-worthiness," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 941–944.

[47] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto, "Understanding factors that affect response rates in twitter," in *Proceedings of the 23rd ACM conference on Hypertext and social media*, 2012, pp. 123–132.

[48] S. N. Firdaus, C. Ding, and A. Sadeghian, "Topic specific emotion detection for retweet prediction," *International Journal of Machine Learning and Cybernetics*, vol. 10, pp. 2071–2083, 2019.

[49] J. Hopcroft, T. Lou, and J. Tang, "Who will follow you back? reciprocal relationship prediction," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1137–1146.

[50] H. S. Kim, Y. J. Suh, E.-m. Kim, E. Chong, H. Hong, B. Song, Y. Ko, and J. S. Choi, "Fact-checking and audience engagement: A study of content analysis and

audience behavioral data of fact-checking coverage from news media," *Digital journalism*, vol. 10, no. 5, pp. 781–800, 2022.

[51] S. Park, J. Y. Park, H. Chin, J.-h. Kang, and M. Cha, "An experimental study to understand user experience and perception bias occurred by fact-checking messages," in *Proceedings of the Web Conference 2021*, 2021, pp. 2769–2780.

[52] M. M. U. Rony, E. Hoque, and N. Hassan, "Claimviz: Visual analytics for identifying and verifying factual claims," in *2020 IEEE Visualization Conference (VIS)*. IEEE, 2020, pp. 246–250.

[53] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[54] J. Chen, Y. Liu, and M. Zou, "User emotion for modeling retweeting behaviors," *Neural Networks*, vol. 96, pp. 11–21, 2017.

[55] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 537–546.

[56] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *Proceedings of the first SIGMM workshop on Social media*, 2009, pp. 3–10.

[57] D. Funke, "This washington post fact check was chosen by a bot," 2018.

[58] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[59] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[60] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[61] E. Brunner and U. Munzel, "The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 42, no. 1, pp. 17–25, 2000.

[62] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[63] C. Buntain, E. McGrath, J. Golbeck, and G. LaFree, "Comparing social media and traditional surveys around the boston marathon bombing." *# Microposts*, vol. 1691, pp. 34–41, 2016.

[64] M. Reveilhac, S. Steinmetz, and D. Morselli, "A systematic literature review of how and whether social media data can complement traditional survey data to study public opinion," *Multimedia tools and applications*, vol. 81, no. 7, pp. 10 107–10 142, 2022.

[65] R. Dolan, J. Conduit, C. Frethey-Bentham, J. Fahy, and S. Goodman, "Social media engagement behavior: A framework for engaging customers through social media content," *European journal of marketing*, vol. 53, no. 10, pp. 2213–2243, 2019.

[66] L. M. Al-Ghamdi, "Towards adopting ai techniques for monitoring social media activities," *Sustainable Engineering and Innovation*, vol. 3, no. 1, pp. 15–22, 2021.

[67] R. Liu, S. Gupta, and P. Patel, "The application of the principles of responsible ai on social media marketing for digital health," *Information Systems Frontiers*, vol. 25, no. 6, pp. 2275–2299, 2023.

[68] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD*, 2017, pp. 1803–1812.

[69] G. W. U. Libraries, "Social feed manager," 2016.

[70] E. Borra and B. Rieder, "Programmed method: Developing a toolset for capturing and analyzing tweets," *Aslib journal of information management*, vol. 66, no. 3, pp. 262–278, 2014.

[71] M. W. Kearney, "rtweet: Collecting and analyzing twitter data," *Journal of open source software*, vol. 4, no. 42, p. 1829, 2019.

[72] C. Byun, H. Lee, and Y. Kim, "Automated twitter data collecting tool for data mining in social network," in *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, 2012, pp. 76–79.

[73] C. Byun, H. Lee, Y. Kim, and K. Ko Kim, "Twitter data collecting tool with rule-based filtering and analysis module," *International Journal of Web Information Systems*, vol. 9, no. 3, pp. 184–203, 2013.

[74] R. Al Bashaireh, M. Zohdy, and V. Sabeeh, "Twitter data collection and extraction: A method and a new dataset, the utd-mi," in *Proceedings of the 2020 the 4th International Conference on Information System and Data Mining*, 2020, pp. 71–76.

[75] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados, "XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 258–266. [Online]. Available: https://aclanthology.org/2022.lrec-1.27

[76] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang *et al.*, "Gpt

(generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, 2024.

[77] L. Belcastro, R. Cantini, F. Marozzo, D. Talia, and P. Trunfio, "Learning political polarization on social media using neural networks," *IEEE Access*, vol. 8, pp. 47 177–47 187, 2020.

[78] S. Y. Lee and Y. Kwon, "Twitter as a place where people meet to make suicide pacts," *Public Health*, vol. 159, pp. 21–26, 2018.

[79] A. Mian and S. Khan, "Coronavirus: the spread of misinformation," *BMC medicine*, vol. 18, pp. 1–2, 2020.

[80] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about covid-19 around the world," *Royal Society open science*, vol. 7, no. 10, p. 201199, 2020.

[81] S. O. Oyeyemi, E. Gabarron, and R. Wynn, "Ebola, twitter, and misinformation: a dangerous combination?" *Bmj*, vol. 349, 2014.

[82] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter," *Cureus*, vol. 12, no. 3, 2020.

[83] S. A. Memon and K. M. Carley, "Characterizing covid-19 misinformation communities using a novel twitter dataset," *arXiv preprint arXiv:2008.00791*, 2020.

[84] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: Detecting covid-19 misinformation in twitter," *Ieee Access*, vol. 8, pp. 155 961–155 970, 2020.

[85] J. S. Brennen, F. M. Simon, P. N. Howard, and R. K. Nielsen, "Types, sources, and claims of covid-19 misinformation," 2020.

[86] Y. Li, T. Grandison, P. Silveyra, A. Douraghy, X. Guan, T. Kieselbach, C. Li, and H. Zhang, "Jennifer for covid-19: An nlp-powered chatbot built for the peopleand by the people to combat misinformation," in *ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID), Seattle, Washington, USA, July 9-10, 2020*, 2020.

[87] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, E. Artemova, E. Tutubalina, and G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. epidemiologia 2, 315–324," 2021.

[88] J. Wei, C. Huang, S. Vosoughi, and J. Wei, "What are people asking about covid-19? a question classification dataset," *arXiv preprint arXiv:2005.12522*, 2020.

[89] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[90] T. G. Coan, C. Boussalis, J. Cook, and M. O. Nanko, "Computer-assisted classification of contrarian claims about climate change," *Scientific Reports*, vol. 11, no. 1, p. 22320, Nov 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-01714-4

[91] B. Dahal, S. A. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Social network analysis and mining*, vol. 9, pp. 1–20, 2019.

[92] A. Aldayel and W. Magdy, "Your stance is exposed! analysing possible factors for stance detection on social media," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, nov 2019. [Online]. Available: https://doi.org/10.1145/3359307

[93] A. Upadhyaya, M. Fisichella, and W. Nejdl, "A multi-task model for sentiment aided stance detection of climate change tweets," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, no. 1, pp. 854–865, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/22194

[94] ——, "Intensity-valued emotions help stance detection of climate change twitter data," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2023, pp. 6246–6254, aI for Good. [Online]. Available: https://doi.org/10.24963/ijcai.2023/693

[95] F. Jost, A. Dale, and S. Schwebel, "How positive is "change" in climate change? a sentiment analysis," *Environmental Science & Policy*, vol. 96, pp. 27–36, 2019.

[96] M. El Barachi, M. AlKhatib, S. Mathew, and F. Oroumchian, "A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change," *Journal of Cleaner Production*, vol. 312, p. 127820, 2021.

[97] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf, "Zephyr: Direct distillation of lm alignment," 2023.

[98] Z. Zhu, Z. Zhang, F. Patel, and C. Li, "Detecting stance of tweets toward truthfulness of factual claims," in *Proceedings of the 2022 Computation+Journalism Symposium*, 2022.

[99] P. Moravec, R. Minas, and A. R. Dennis, "Fake news on social media: People believe what they want to believe when it makes no sense at all," *Kelley School of Business research paper*, no. 18-87, 2018.

[100] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang, "Retweet prediction with attention-based deep neural network," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 75–84.

[101] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet prediction based on topic, emotion and personality," *Online Social Networks and Media*, vol. 25, p. 100165, 2021.

[102] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 2335–2338.

[103] H. Yu, X. F. Bai, C. Huang, and H. Qi, "Prediction of users retweet times in social network," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 5, pp. 315–322, 2015.

[104] D. Huang, J. Zhou, D. Mu, and F. Yang, "Retweet behavior prediction in twitter," in *2014 Seventh international symposium on computational intelligence and design*, vol. 2. IEEE, 2014, pp. 30–33.

[105] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2012, pp. 46–50.

[106] S. N. Firdaus, C. Ding, and A. Sadeghian, "Retweet prediction considering user's difference as an author and retweeter," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 852–859.

[107] J. Zhang, J. Tang, J. Li, Y. Liu, and C. Xing, "Who influenced you? predicting retweet via social influence locality," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 9, no. 3, pp. 1–26, 2015.

[108] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[109] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[110] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, "Topic-conditioned novelty detection," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 688–693.

[111] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[112] R. Pfitzner, A. Garas, and F. Schweitzer, "Emotional divergence influences information spreading in twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, no. 1, 2012, pp. 543–546.

[113] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[114] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky, "Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter," in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 2023, pp. 5597–5607.

[115] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[116] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.