A USER-BASED STANCE ANALYSIS FOR GAUGING PUBLIC OPINION WITH STANCE DETECTION IN TWITTER DATA

by

Ali Almadan

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2022

Approved by:

Dr. Mary Lou Maher

Dr. Albert Park

Dr. Wenwen Dou

Dr. Frederico Batista Pereira

Dr. Jason Windett

ABSTRACT

ALI ALMADAN. A User-Based Stance Analysis for Gauging Public Opinion with Stance Detection in Twitter Data. (Under the direction of DR. MARY LOU MAHER)

Stance detection in social media data has received attention in recent years as an approach to determine the standpoint of users toward a target of interest, such as a person or a topic included in Twitter data. Although interviewing, surveying, and polling representative populations have long proven reliable methods for analyzing public opinion, these methods suffer from various limitations, including high costs and an inability to be collected retrospectively. On the other hand, detecting and analyzing social media trends through natural language processing approaches, such as text classification, offers a valuable alternative or complementary approach to gathering, analyzing, monitoring, and understanding public opinion on emerging issues.

Existing stance detection and analysis studies use multiple methodologies and strategies to determine and analyze the standpoint of Twitter users towards a target. These techniques feature strengths and weaknesses, and the literature lacks studies investigating the broad implications of using such methods for public stance measurements. Understanding these implications is crucial to the validity, interpretation, and replicability of research findings.

In this dissertation, we first introduce the concept of user-based stance analysis and highlight the difference between user-based and tweet-based stance analyses. We describe the relevance of user-based stance analysis to the measurement of public opinion. We suggest that the stance of Twitter users, instead of the stance presented in a tweet's content, must be the core aspect of stance analysis for measuring public opinion. Therefore, we claim that a user-based stance analysis is more aligned with the concept of public opinion than a tweet-based stance analysis. Second, we compare the results of measuring public opinion with tweet-based and user-based stance analyses

from Twitter data and demonstrate that each produces statistically different results. Third, we present findings that, while a tweet-based stance analysis is sensitive to the presence of social bots, a user-based stance analysis provides a more robust measure of public opinion with minimal impact from social bots. Fourth, we describe the design and evaluation of StanceDash, a web-based dashboard that assists end users measure, analyze, and monitor public opinion through a user-based stance analysis of Twitter data.

DEDICATION

This dissertation is dedicated to:

The memory of my beloved father, Abdulgani Almadan, who passed away during the second year of my Ph.D. journey on 06/30/2019.

and

My precious mother, Wedad Alzayer, who never stopped keeping me in her prayers and thoughts while pursuing my dream abroad.

and

The rest of my family and friends. Without their endless support, encouragement, and patience, none of this research would have been possible.

ACKNOWLEDGEMENTS

First, I would like to express my greatest gratitude to my family: My father (may his soul rest in peace), mother, siblings, and my wife for their endless support and encouragement while pursuing my doctorate degree at The University of North Carolina at Charlotte. I also would like to thank my friends and colleagues, especially Hamad Alsaleh, in the Ph.D. program for their support and fruitful discussions during this joyful journey.

I also extend my gratitude to my advisor, Dr. Mary Lou Maher, for her guidance and support since the first day of my doctorate degree at UNC Charlotte. Special thanks to the members of my dissertation committee: Dr. Albert Park, Dr. Wenwen Dou, Dr. Frederico Batista Pereira, and Dr. Jason Windett for their guidance over the past 5 years.

Finally, I am grateful to the Saudi Arabian Government and the Ministry of Education in the Kingdom of Saudi Arabia for sponsoring my graduate studies in the United States.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API An acronym Application Programming Interface

AWS An acronym Amazon Web Services

BERT An acronym for Bidirectional Encoder Representations from Transformers

GPU An acronym for Graphics Processing Unit

K-S An acronym for Kolmogorov-Smirnov Normality Test

LDA An acronym for Latent Dirichlet Allocation

LSTM An acronym for Long Short-Term Memory

NB An acronym for Naive Bayes

NLP An acronym for Natural Language Processing

NLTK An acronym for Natural Language Toolkit

REST An acronym for Representational State Transfer

SD An acronym for standard deviation

SVM An acronym for Support Vector Machine

WHO An acronym for World Health Organization

CHAPTER 1: INTRODUCTION

"The core advantage of data is that it tells you something about the world that you didn't know before."

Hilary Mason

In recent years, the evolution of the Internet has made information easily and widely accessible. Social media platforms in general and Twitter, in particular, are fast-growing microblogging systems that have reshaped online communication and how people connect to each other. With unrestricted access, such platforms have enabled researchers to study a variety of phenomena and understand public opinion and behavior from online posts as users share their thoughts and express their opinion and viewpoints about different topics. On Twitter alone, there are more than 363 million active users and more than 10,000 tweets sent per second[1]. Due to the large and overwhelming content on social media, computational tools were needed to detect, analyze, visualize, and understand how people express their opinions on social media. This includes discovering underlying discussion topics, analyzing emotions [2, 3], and predicting real world outcomes [4, 5, 6]. There are several research problems that emerged from text data from social media. Some of these research problems are sentiment analysis (also known as opinion mining), emotion detection and recognition, argument mining, sarcasm/irony detection, veracity detection, and fake news detection.

One recent problem that has emerged recently from the valuable social media data is

[1] https://www.internetlivestats.com/twitter-statistics/

stance detection. Stance detection is the problem of classifying the attitude expressed toward a target of interest where the input is a text and the output is a stance label [7]. Stance detection has different social and political applications, such as polling / surveillance and rumor detection [8, 9, 10]. In addition, stance detection can be seen as a method of understanding public opinion from Twitter data.

Public opinion plays a significant role in democratic societies. Generally speaking, public opinion is the measure of the public's viewpoint about a particular topic, issue, party, or individual political figure. For example, public opinion can gauge the attitude of the general public toward abortion legalization in the United States. One of the early definitions of public opinion was the definition of Floyd H. Allport in 1937 [11]. In his article *Toward a Science of Public Opinion*, Allport describes public opinion as follows:

> The term public opinion is given its meaning with reference to a multi-individual situation in which individuals are expressing themselves, or can be called upon to express themselves, as favoring or supporting (or else disfavoring and opposing) some definite condition, person, or proposal of widespread importance, in such a proportion of number, intensity, and constancy as to give rise to the probability of affecting action, directly or indirectly, toward the object concerned.

In fact, public opinion is the backbone of a democratic society. Public opinion informs decision making at both the individual and the government levels. At the individual level, for example, public opinion determines which person gets elected. At the government level, public opinion determines which policies are implemented or discontinued.

Traditionally, interviews and surveys have been the most common instruments for measuring public opinion. Although self-report through surveys and interviews is a

well-known method in social and behavioral sciences, it suffers from major limitations, such as sampling and social desirability biases. The limitations of self-report methods could affect the validity of the measurements. In addition, these traditional methods tend to be financially costly and time-consuming. These features of self-report through surveys and interviews limit the extent to which the collected data can inform about time trends. More importantly, data usually cannot be collected retrospectively in a reliable manner using such instruments. In contrast, using stance detection in widely accessible and available social media posts that convey opinions about targets of interest can provide large amounts of public opinion data observed over longer periods of time.

Many studies have focused on detecting the stance in Twitter data by training and evaluating machine learning classifiers. In recent years, studies have shown the impact of different classification approaches and the extraction of textual features on improving the performance of stance detection [12, 13, 14]. However, studies in the literature do not investigate the impact of critical methodological issues, decisions, and alternatives on measuring public opinion with stance detection in Twitter data. Rather than focusing solely on improving the performance of stance detection, we examine how stance detection and analysis can be used effectively to measure public opinion from Twitter data. Ultimately, the results in this dissertation provide a methodology to apply stance detection in Twitter data to understand, analyze, and monitor public opinion from Twitter data.

1.1 Research Motivation

Traditional measures of public opinion, such as scientific polls and interviews, have been effective instruments for measuring public opinion. However, they suffer from disadvantages and limitations. First, a traditional public opinion study requires the recruitment of participants. By itself, the recruitment process can suffer from biases such as sampling bias [15, 16]. Second, these measures are prone to self-report biases

such as social desirability bias [17, 18], where respondents tend to provide responses that are favorable to others [19]. Third, scientific polling and interviews have limited scalability. Fourth, it is challenging to use traditional methods to gauge public opinion as events emerge in real time. Finally, data from these traditional instruments cannot be collected retrospectively. In general, traditional methods are human intensive. As a result, they are expensive and time consuming.

The limitations of traditional methods for gauging public opinion and the progress in stance detection research motivated the current research. The need to find alternatives that overcome the disadvantages and limitations is high. Compared to other natural language processing tasks, such as sentiment analysis [20] and emotion detection [21], stance detection is the most related task to public opinion, by definition. Using stance detection in Twitter data overcomes the limitations of surveys. However, the literature lacks studies that examine how stance detection can be used to effectively measure public opinion. This dissertation proposes user-based stance analysis for robust measurement of public opinion from stance detection in Twitter data. The research presented in this dissertation benefits the research in the computer and social sciences fields.

1.2 Thesis Statement and Research Questions

Stance analysis is a way to measure public opinion with stance detection in Twitter data. Although stance analysis can be either tweet-based or user-based, the literature lacks studies that show how stance analysis can be effectively used to measure public opinion from Twitter data. In this dissertation, we have the following thesis statement:

Thesis Statement: A user-based stance analysis to measure public opinion in Twitter data produces more meaningful and statistically different results from a tweet-based analysis, mitigates the impact of social bots, and helps to monitor and understand changes in public opinion.

Broadly speaking, this research addresses the thesis by pursuing an answer to the

general question: How can stance detection be used effectively to measure public opinion from Twitter data? Our specific research questions are as follows:

- **RQ1)** How does the report of public opinion from Twitter data differ when based on the analysis of tweet-based stance vs user-based stance in Twitter data?
- **RQ2)** How does a user-based stance analysis mitigate the impact of social bots on measured public opinion in Twitter data?
- **RQ3)** How does visualizing a user-based stance in a dashboard assist end users monitor, analyze, and understand public opinion in Twitter data?

1.3 Contributions

This dissertation makes several contributions to the literature. There contributions are:

1. We introduce the concept of user-based stance analysis to measure public opinion with stance detection in Twitter data.
2. For the first time, we provide a comparison between two analysis methods of using stance detection to measure public opinion from detection: tweet-based and user-based stance analyses.
3. We examine the effectiveness of a user-based stance analysis in mitigating the impact of social bots on measuring public opinion with stance detection in Twitter data.
4. We introduce the design evaluation of a dashboard that is used to monitor and analyze public opinion in Twitter data by employing a user-based stance analysis and other Natural Language Processing techniques such as topic modeling.

1.4 Dissertation Organization

This dissertation is organized as follows: Chapter 1 highlights the research problem, motivation, and research questions. Chapter 2 provides a background and literature review related to public opinion, public opinion in Twitter data, and bot manipulation of public opinion in Twitter data. Chapter 3 describes the data collection. Chapter 4 presents a statistical comparison between tweet-based and user-based stance analysis in Twitter data. Chapter 5 shows how a user-based stance analysis mitigates the impact of social bots on measuring public opinion from Twitter data. Chapter 6 describes the design and evaluation of a web-based dashboard to measure, analyze, and monitor public opinion in Twitter data. Finally, Chapter 7 outlines the limitations, future work directions, and the conclusion.

CHAPTER 2: BACKGROUND ON PUBLIC OPINION AND STANCE DETECTION

This chapter reviews the background for the current dissertation research. In particular, it discusses the background for public opinion, stance, and stance detection in Twitter data.

2.1 Public Opinion

Broadly speaking, the concept of public opinion can be seen as any collection of individual opinions designated [22]. Over 100 years of scholars, journalists, statisticians, and pollsters have transformed the understanding of collective opinion in areas such as individual opinion formation, group opinions, elite influence on public opinion, the connection of public opinion and government responsiveness, as well as issue-specific opinion dynamics[1].

Proper measurement of public opinion is a critical component in evaluations of democratic responsiveness in public institutions. Advancements in survey techniques and methodologies, as well as interviews with citizens, allowed for a wide-scale analysis of opinion formation, belief structures, and what influences opinion change. [24]. However, these more advanced techniques reinforced earlier theories that most individual Americans were and remain fairly uninformed [25].

The scholarship in mass public opinion found that the macro level, or aggregation of individuals, provided an informed, responsive public "mood" that would change over time and in response to events [26]. Furthermore, government institutions respond to these mood changes by adjusting the ideological intensity of the outcomes [27]. Although these initial studies look at aggregation in a single ideological space, and not

[1]See [23] for a detailed treatment.

across issue dimensions, recent innovations have allowed for more naunced measures of public mood across different issue areas that will greatly improve our understanding of mass public opinion and government response [23].

Although the overwhelming majority of public opinion research uses traditional survey techniques such as online polling, in-person interviews, and telephone banks, the expansion of the use of social media by both the general public and officials allows for a new avenue of research. Using social media platforms as a new mode to measure public preferences has become increasingly widely used in a variety of political contexts. For example, [28] used Twitter during the German National Elections in 2009 to accurately predict party positions, public sentiment, and election results. In one study [29], researchers used Twitter data to predict the Brexit election results using nearly 23 million unique tweets. In another study [30], researchers used polling from the 2016 US Presidential Election to validate Twitter opinion trends that detect shifts in opinion earlier than traditional public opinion polling techniques.

2.2 Stance and Stance Detection

This section highlights background related to stance and stance detection. The following subsections explain the concept of *stance* from a linguistic point of view, stance detection and related NLP tasks, stance detection algorithms, and data.

2.2.1 Stance in Linguistics

From a linguistic perspective, there are a few definitions of *stance* in the literature. An early definition of stance is the definition by Biber and Finegan in 1988 [31]. Biber and Finegan define stance as follows:

> *Lexical and grammatical expression of attitudes, feeling, judgments, or commitment concerning the proportional content of the message.*

In this definition, Biber and Finegan outline three components of stance-taking: 1) the attitude of a person, 2) the feelings and emotions of the person, and 3) the

judgment of the person.

Another definition of stance is the definition proposed by Du Bois in 2007 [1]. In the *The Stance Triangle* article, Du Bois defines a stance as follows:

> *A public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and other), and aligning with other subjects, with respect to any salient dimension of the sociocultural field.*

In his definition, Du Bois outlines three pillars of stance-taking in discourse: evaluation, positioning, and alignment. As shown in Figure 2.1, someone (subject1) evaluates an object, positions themselves and aligns with others (subject 2).

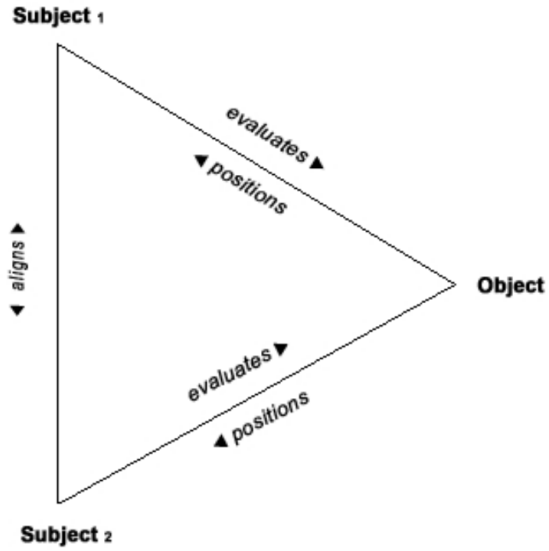

Figure 2.1: Du Bois Stance Triangle [1]

To illustrate the idea of the stance triangle, consider the following simple statement that someone posted on Twitter in mid-2020:

> *I am sad that we do not have a vaccine for COVID-19 yet*

In this simple statement, the speaker (subject 1) evaluates an object (vaccinations) and positions himself regarding the object.

It is clear from the definitions of stance [31, 1] that stance-taking is a complex process that includes more than sentiment and emotions. Therefore, the stance of a person toward a subject cannot be detected simply by detecting the emotion or polarity in a text.

2.2.2 Stance Detection and Related NLP Tasks

Stance detection in social media data and in Twitter data, in particular, has received considerable attention for the stance detection task over the past few years, possibly due to the availability of the data. Stance detection is the problem of automatically determining the standpoint expressed in a tweet toward a target of interest, such as a person or an issue [32, 33]. Stance detection is also referred to as stance identification [34], stance classification [35], and stance prediction [36, 37]. Although the selection of the stance labels depends on the stance target, the commonly used stance labels are favor, against, and neutral. The goal of stance detection is to identify how a user positions themselves (explicitly or implicitly) with regard to a target, such as a presidential candidate [38, 33, 32]. Moreover, stance detection can be performed to determine the standpoint toward either a single target or multiple targets. Stance detection in social media data has numerous applications, such as the detection of fake news, public health surveillance, and information retrieval [33]. For example, [39] used deep learning stance detection to detect fake news. The simplest form of stance detection is single-target stance detection, which we can formulate as follows:

$$Stance(T, U|G) = \{Favor, Against, Neutral\} \tag{2.1}$$

Where the input is a text T and user U, and given a target G, the user's stance can be assigned any of the labels Favor, Against, or Neutral [32].

The selection of the stance labels depends on the problem in the research. Although researchers commonly use favor, against, and neutral labels in the literature, other

researchers have used other labels. For example, the study by Gorrell et al. (2019) [40] used the labels Support, Deny, Query, Comment to detect the stance of Twitter users towards mental disorders. Other studies used the stance labels neither, none, and unclear instead of neutral [41, 42, 43].

In recent years, various NLP tasks have been used to measure public opinion in Twitter data. Some of these NLP tasks are sentiment analysis (also known as opinion mining) [20], and emotion detection (also known as emotion recognition) [44]. Stance detection is different from sentiment analysis, a natural language processing approach that is used for opinion mining [20]. Sentiment analysis, in its most basic form, relies on the polarity of the text in tweets to determine whether the text conveys positive, negative or neutral sentiment. Sentiment analysis can be formulated as shown in Equation 2.2.

$$Sentiment(T) = \{positive, negative, neutral\} \tag{2.2}$$

Where T is the *text* of the tweet.

Similarly, stance detection is different from emotion detection in textual data. Emotion detection is a branch of analysis that deals with extracting the state of mind of the user [45]. In fact, sentiment is defined as the effect of emotions [46]. As an example, emotion detection can be formulated as follows for Ekman's six basic emotions [47]:

$$Emotion(T) = \{Happiness, Sadness, Fear, Disgust, Anger, Surprise\} \tag{2.3}$$

From Equations 2.1, 2.2, and 2.3, it is clear that stance detection, sentiment analysis, and emotion detection are distinct tasks. To illustrate the difference, Figure 2.2 shows the different results of sentiment analysis, emotion detection, and stance detection on a sample tweet.

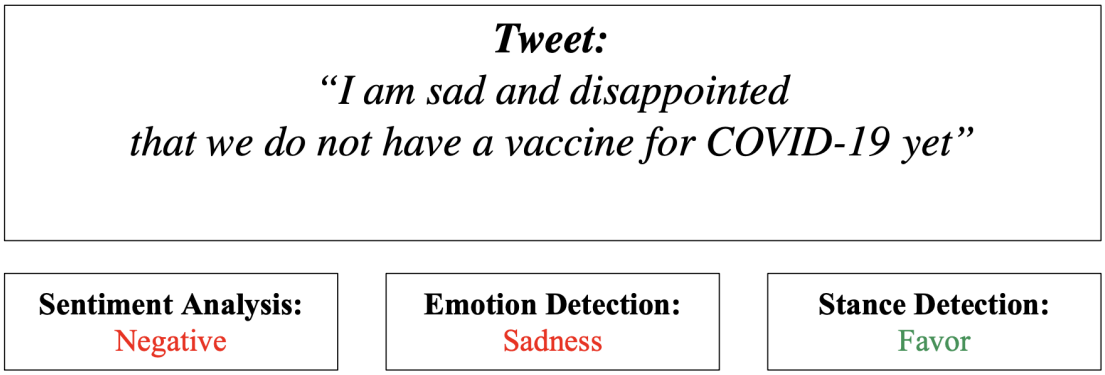

Figure 2.2: An example to highlight the difference between sentiment analysis, emotion detection, and stance detection

From the example in Figure 2.2, the three tasks provided different results. On the one hand, sentiment analysis has determined that the sentiment of the text was negative, perhaps due to the presence of the words "sad" and "disappointed". Similarly, for emotion detection, the emotion conveyed in the text is sadness, as the tweet user explicitly stated that they were sad. On the other hand, it is clear that the Twitter user who posted the tweet is in support of COVID-19 vaccination.

Although stance detection is more relevant to public opinion than sentiment analysis and emotion detection, limited studies have used stance detection to measure public opinion toward targets of interest. In one study, researchers used stance detection in Twitter data to measure public opinion toward the Job Creation Bill in Indonesia [48]. Another study by researchers investigated public opinion on e-cigarettes in Indonesia [49]. In general, the literature lacks studies exploring how stance detection can be used as a robust measure of public opinion in Twitter data.

2.2.3 Stance Detection Algorithms

There are several approaches and algorithms for stance detection in Twitter data. The most common approach for stance detection in Twitter data leverages supervised machine learning to classify the stance into one of the stance categories. In a traditional supervised learning approach, tweets are collected and then annotated by

human annotators, and the data is used to train and test the classifier. Although the size of the datasets used to train and test the classifier varies in the literature, the size ranges from a few hundred tweets to tens of thousands of tweets. After human annotators annotate the tweets, NLP preprocessing steps are applied to the tweets, if necessary. Then, the annotated dataset is split into training and testing sets. Once a classifier achieves high accuracy, it is applied to the data, and a stance analysis is carried out. The general supervised stance detection steps are shown in Figure 2.3.

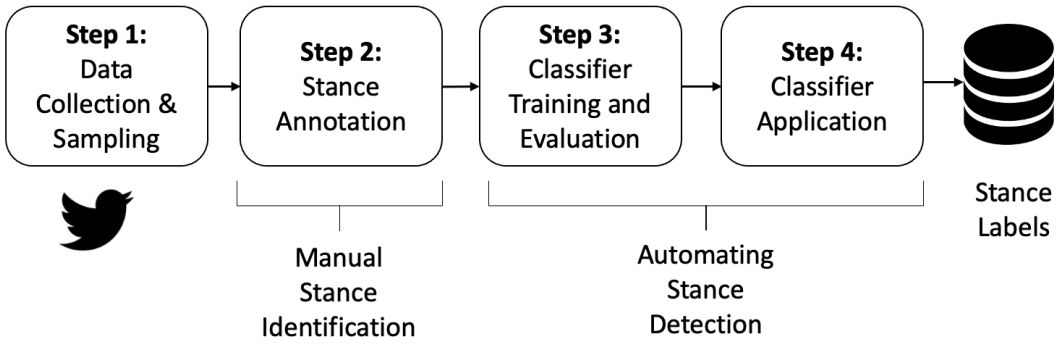

Figure 2.3: General supervised stance detection steps in the literature

In Step 1, data are collected from Twitter using selected keywords or hashtags (or both). Once the data are collected, human annotators read and determine the stance of the tweet (favor, against, or neutral), as shown in Step 2. Then, the annotated tweets can be used to train, validate, and test a supervised machine learning text classifier (Step 3). Once a classifier with reasonable performance is implemented, it can be applied to new unseen tweets to detect the stance of the tweets (Step 4). Finally, the stance labels can be used to analyze the stance toward an issue of interest from Twitter data.

Stance *detection* can be either at the tweet level or at the user level. At the tweet level, the content of the tweet is used to classify the stance of the tweet. On the other hand, stance detection at the user level incorporates user-related features, such as network interactions, to classify the user's stance. To measure public opinion with

stance detection from Twitter data, the stance of the Twitter user must be the center of the analysis rather than the tweet. While most studies in stance detection performed stance detection at the tweet level, a few studies followed a different direction in detecting the stance at the user level. In one study [12], researchers showed that the use of network features can improve the task of automatically detecting the stance of Twitter users using a supervised classification approach. In another study [50], researchers used an unsupervised learning approach to detect the stance of Twitter users based only on a few of their tweets. However, one limitation of the two studies is that they used datasets that covered short periods of time. For example, the dataset in [50] covered approximately one week of data for each stance target. Therefore, the authors assigned one single stance to each user under the assumption that a user's stance does not change over a short period of time. Although these approaches have been shown to improve the performance of the stance detection task, a user's stance can change over longer periods of time. Therefore, a single user can have multiple stances at different points in time, as shown by [51] in their analysis of the misalignment between the stance expressed by Twitter's users and their responses in surveys. To understand the change in public opinion and public opinion's dynamics, it is essential to incorporate temporal analysis and explore how users change their stance over time.

A wide range of classification approaches have been used to detect the stance in tweets. Stance detection approaches include classical machine learning algorithms and deep learning, as shown in recent surveys [33, 32]. In classical machine learning approaches, support vector machine (SVM) is the most popular algorithm for stance detection. Other algorithms include logistic regression and Naive Bayes (NB). Popular deep learning approaches include Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory (LSTM).

2.2.4 Stance Detection Data

While few studies have studied stance detection on Facebook posts [52, 53], our observation is that the focus in the literature is on Twitter as the primary platform for stance detection. This is possibly due to the availability of the data and the ease of using the Twitter API endpoints. According to Twitter's documentation, there are two main approaches for collecting Twitter data. These approaches are as follows.

- **Twitter Streaming API:** provides real-time data collection and allows tracking of specific keywords and hashtags.
- **Twitter REST API:** Follows a client-server architecture where a request is sent to the server and the results are sent back to the client. It provides a tool to search for tweets using keywords and hashtags. The public REST API has the standard API and the academic API. While the REST API allows researchers to retrieve tweets posted in the last seven days only, the academic API allows access to historical Twitter data regardless of when the tweet was posted.

Each of the streaming and REST APIs have different versions. For example, the streaming API can provide a 1% sample of tweets instead of streaming all tweets. On the other hand, the basic REST API allows searching and retrieving tweets in the last seven days, while the academic API allows searching and retrieving all tweets with a limitation of 10 million tweets per month.

Each of the streaming and REST APIs has strengths and weaknesses. For example, while the streaming API provides a utility to collect tweets as users tweet in real-time, the REST API does not provide a way to search for and retrieve deleted tweets. A tweet can be deleted by the user or Twitter for violating their terms of service. One of the limitations of the streaming API is that the API cannot return more than 1% of the overall volume of Twitter tweets at a single time point. When a query's results exceed 1% of the overall volume, the API will return a sample of the tweets that

satisfy the query. For both APIs, the input argument to retrieve the tweet can be a keyword or a hashtag.

Several Twitter annotated datasets are available for the task of stance detection. The available English datasets in the literature, as summarized in Table 2.1, were collected using the two approaches described in this section. We acknowledge that some studies annotated more datasets using the two approaches, but the datasets are not made available by the authors or are not in English [54, 55, 56].

A challenge with the current datasets is that Twitter does not allow researchers to publish the text of the tweets, but only tweet identifiers can be published. Therefore, hydrating all tweets might not be possible for various reasons such as the deletion of the tweets and/or the suspension of the Twitter accounts. As a result, the public stance distribution could change when the data collection methods change.

Table 2.1: Currently available stance detection datasets

Authors	**Target**	**Stance Labels**	**Size**	**Collection Method**	**Collection Argument**
Mohammad et al. (2016) [41]	Atheism, Climate, Feminism, Hillary, Abortion	Favor, Against, Neither	4,870 tweets	REST API	hashtags
Sobhani et al. (2017) [57]	Clinton, Sanders, Trump, Cruz	Favor, Against, Neither	4,455 tweets	REST API	keywords
Addawood et al. (2017) [42]	Individual privacy	Favor, Against, Neither	3,000 tweets	Streaming API	keywords
Aker et al. (2017) [58]	Mental Disorder	Support, Deny, Query, Comment	401 tweets	Streaming API	hashtags & keywords
Gorrell et al. (2019) [40]	Rumors	Support, Deny, Query, Comment	6,634 tweets	Unclear	Unclear
Grcar et al. (2017) [59]	Brexit	Leave, Remain, Neutral	35,000 tweets	Unclear	hashtags
Conforti et al. (2020) [60]	Merge and acquisition of companies	Support, Refute, Comment, Unrelated	51,284 tweets	REST API	keywords
Cotfas et al. (2021) [61]	COVID-19 vaccination	Favor, Against, Neither	7,530 tweets	REST API	keywords
Almadan et al. (2022) [43]	COVID-19 vaccination	Favor, Against, Unclear	800 tweets	Streaming API	Hashtags & Keywords

CHAPTER 3: DATA COLLECTION

In this chapter, we discuss how we collected Twitter data for the dissertation studies and analyses. First, we describe the software tools, libraries, and services that we used to collect and classify the data. Next, we describe the methodology for collecting and storing tweets.

3.1 Tools, Libraries, and Services

In this section, we explain in detail the tools, libraries, and services that we used to collect the data. We used Python for data collection, preprocessing, and classification. We also relied on previously developed tools and libraries that were compatible with Python. Tools and libraries include:

- **Python 3.7:** A programming language that allows data processing, manipulation, and modeling.
- **Tweepy:** A library in python that is used to retrieve tweets. A valid Twitter developer account authentication is required.
- **NLTK Natural Language Processing Toolkit:** A library to process text for natural language processing in Python.
- **Google Colab Pro+:** A paid Jupyter notebook environment that runs entirely in the cloud and allows the use of GPUs for computationally expensive operations such as training neural networks.
- **ScikitLearn:** A library in Python that provides a variety of supervised and unsupervised machine learning algorithms in Python.

- **SciPy:** A library in Python that provides statistical tests such as correlation and the K-S test.
- **TensorFlow:** A library in Python for creating machine learning applications and widely used for neural networks in Python.
- **Pandas**: An open-source library for data manipulation for machine learning tasks in Python.
- **Matplotlib**: An open-source library for data visualization in Python.
- **Gensim:** An open-source library for Natural Language Processing tasks and modeling.

3.2 Twitter Data Collection

The novel coronavirus disease 2019 (COVID-19) was originally observed as a cluster of unexplained pneumonia cases in Wuhan, China, at the end of 2019, and on March 11, 2020, the World Health Organization (WHO) declared COVID-19 a global pandemic [62]. By the end of March 2021, there were more than 130 million cases and 2.8 million deaths worldwide. It was widely believed that COVID-19 vaccination was essential to protect the vulnerable population and return to normalcy. Vaccination has been associated with the successful eradication and prevention of previous infectious diseases such as measles and polio [63]. According to the World Health Organization[1], herd immunity requires 95% and 80% of the population to be vaccinated against measles and polio, respectively. Although the threshold for COVID-19 was not yet known, it was accepted that herd immunity could only be achieved when a large percentage of the population receives the vaccine. Despite the critical role of COVID-19 vaccination in ending the pandemic, numerous studies suggested that vaccine hesitancy existed and was expected to hinder COVID-19 vaccination efforts [64, 65, 66, 67].

[1]https://www.who.int/news-room/q-a-detail/herd-immunity-lockdowns-and-covid-19

Due to the importance of the vaccination topic and the considerable attention it received after COVID-19, we collected tweets related to vaccination using Twitter's streaming API between June 2 and November 28, 2021. During this period, some COVID-19 vaccines, such as the Pfizer and Moderna vaccines, were widely available to the population and people used Twitter to express their opinions about the vaccine. We favored the streaming method over the archive search method for data collection because we anticipated that a considerable number of tweets were likely to violate Twitter. The social media company said in a statement published on its website that it was going to suspend accounts that repeatedly shared misleading information.

Although there were several Twitter English datasets related to COVID-19 [68, 69, 70], we decided to collect our own using Twitter's Streaming API because other datasets require hydrating tweets (refer to Section 2.2.4 for more details). Tweets cannot be hydrated if they were deleted or the users who posted them were suspended. Collecting our data enabled us to capture tweets in real-time as they were being posted and before possible deletion. In addition, other datasets were collected using keywords related to coronavirus and were general to the COVID-19 pandemic but not to vaccination. We wanted to capture the stance towards vaccination in the era of COVID-19.

To ensure uninterrupted and reliable data streaming, we set up an account on Amazon Web Services (AWS). AWS is a cloud-computing platform developed by Amazon for individuals and businesses and provides high-performance CPUs and storage. We used Amazon Elastic Computing (EC2) with 16GB memory to connect and stream Twitter data directly from Twitter to Amazon General Purpose SSDs. The data was streamed into files of 100,000 tweets in size rather than storing each tweet in a single file to facilitate the movement, reading, and processing of the data. When a file reached 100,000 tweets, the file was closed, a new file was created, and the tweets were streamed into the new file. The data stream and storage process is

shown in Figure 3.1.

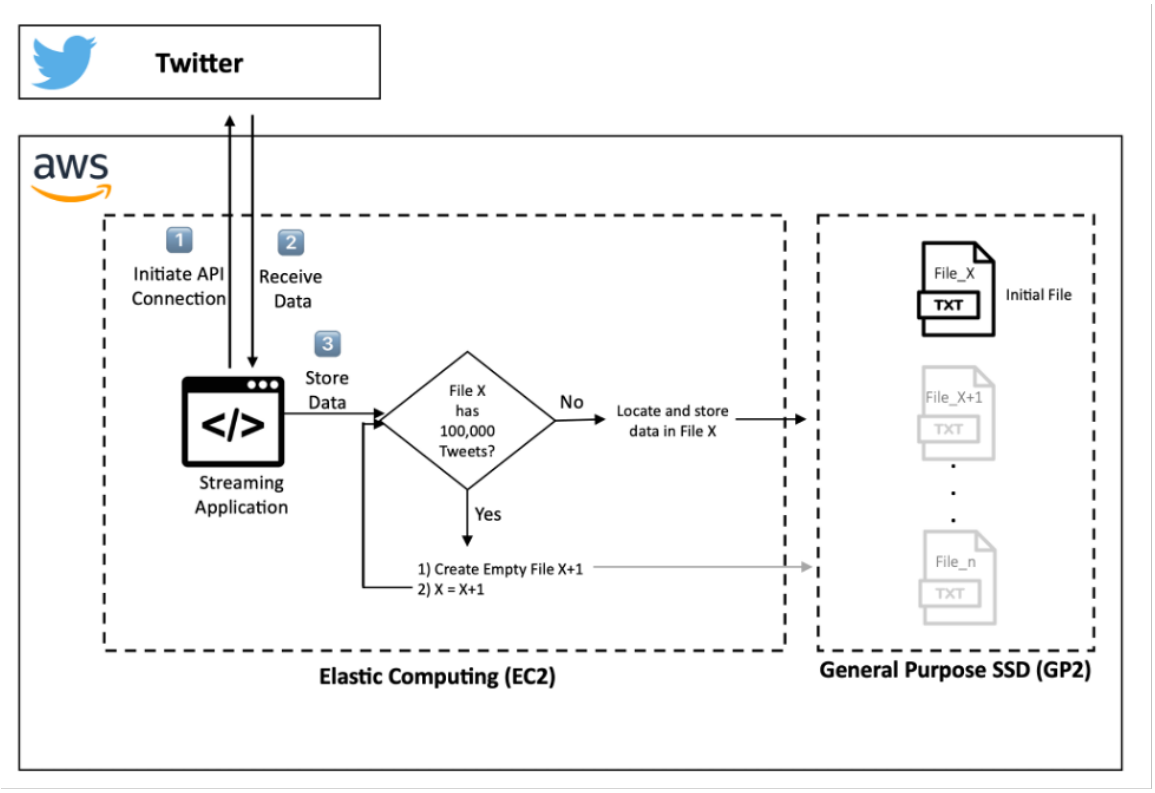

Figure 3.1: Data collection using Amazon Web Services AWS

To filter the stream, we used general vaccine-related keywords and hashtags from the literature to examine pro- and anti-vaccination (ProVax and AntiVax) users on Twitter [71, 72]. For example, the hashtag #vaccinessavelives conveys a favor stance toward vaccination. On the other hand, the hashtag #vaccineskill conveys an against stance. The complete list of keywords and hashtags is presented in Table 3.1. We excluded retweets and tweets with external links from the data collection. Tweets that were not in English were also excluded from the data collection stream. The final dataset had 24,806,152 tweets and 4,664,146 unique users. Figure 3.2 shows the volume of tweets collected per day. Figure 3.3 shows the 10 most common geographic locations in the data within the United States.

Because the data were streamed and stored in JSON objects with all attributes that were associated with the users and tweets, we needed to extract the specific

Table 3.1: List of keywords and hashtags that were used to filter the stream.

Keywords	'vaccine','vaccines','vaccination', 'vaccinations'
Hashtags	'#antivaxxers', '#antivax', '#antivaxxer', '#vaccineagenda', '#vaccineswork', '#novaccineforme', '#antivaccination', '#novaccinemandate', '#vaccinesprotect', ' #vaccines4results', '#vaccinesaregenocide', '#novaccineforme', ' #vacccinessavelives', '#vaxwithme', '#vaccineinjury', '#vaccinedeath', '#vaccinedamage', '#novaccine', '#vaccinefraud', '#vaccineskill', '#vaccinesarepoison

Table 3.2: The user and tweet attributes that were extracted from the JSON objects.

Attribute	Entity	Description
created_at	tweet	The date and time when the tweet was posted
place	tweet	The geographic location Twitter assigns to the tweet
text	tweet	The content of the tweet. If the tweet is long, it will be truncated
text	extended_tweet	The full content of the tweet if the tweet was long and truncated
id	tweet	The unique identifier of the tweet
screen_name	user	The display name of the user who posted the tweet
id	user	The unique identified of the user who posted the tweet

attributes that were needed for this research. To achieve this, we developed a Python program that read the JSON objects, extracted the needed attributes, and stored the new data in Excel format (xlsx). The attributes we extracted are shown in Table 3.2.

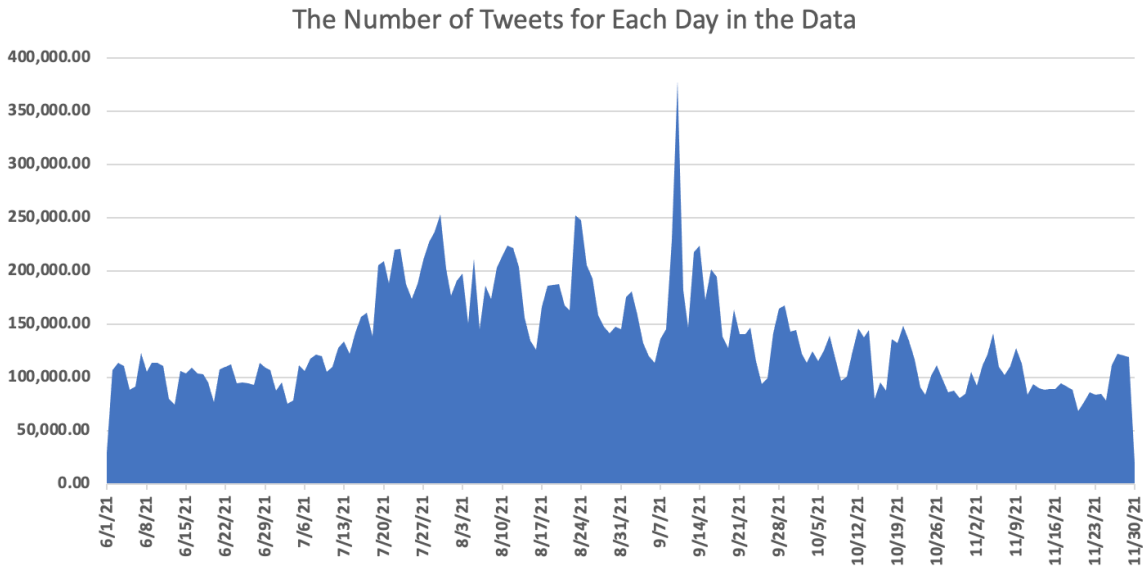

Figure 3.2: The volume of tweets per day

Location	Frequency
Los Angeles, CA	5784
Manhattan, NY	4475
Florida, USA	4034
Chicago, IL	3486
Houston, TX	2670
Georgia, USA	2661
Brooklyn, NY	2458
Pennsylvania, USA	2376
Texas, USA	2369
Washington, DC	2279

Figure 3.3: Most common locations in the data

CHAPTER 4: A STATISTICAL ANALYSIS OF THE DIFFERENCE BETWEEN TWEET-BASED AND USER-BASED STANCE IN TWITTER

4.1 Introduction to Measuring Public Opinion with Stance Detection

A central tenet of representation in democratic systems is a connection between public preferences and policy outcomes offered by elected officials. Historically, the public's opinion is measured through survey instruments administered by newspapers, political organizations, non-profit groups, and many others. Although a great deal of understanding the public's behavior has come from these methodologies, traditional public opinion polling is beginning to suffer high levels of non-response rates, increased costs associated with collecting the data, and an overreliance on non-representative samples with survey weights that potentially bias the outcome. As such, alternative methods of gauging the pulse of the public can better allow individuals and elected officials to respond more appropriately and with greater speed. An increasingly popular alternative is to use stance detection. In this manuscript, we show how macrobehavior and opinions toward vaccination in the era of the COVID-19 pandemic can be measured using social media data and stance detection methods.

Stance detection has received considerable attention in recent years, especially after the SemEval competition in 2016. In SemEval 2016 Shared Task 6, training and evaluation tweet datasets were provided to predict whether the tweets were in favor, against or did not have a position toward the five targets: Atheism, Hillary Clinton, Climate Change is a real concern, the feminist movement, and abortion [41]. Applying stance detection on widely available and accessible Twitter data and analyzing the results provide an alternative to long-used traditional methods, such as surveys and interviews, to measure public opinion as users turn to social networks to express their

opinions. In other cases, public opinion obtained from stance detection in Twitter data can be seen to complement the public opinion measured by traditional methods [51]. Stance in Twitter data can be formulated as follows:

$$Stance(T, U|G) = \{favor, against, neutral\} \tag{4.1}$$

Where T is the *text* of the tweet, U is the *user* who posted the tweet, and G is the *target* of the stance [32].

Stance detection is different from sentiment analysis, a natural language processing approach that is used for opinion mining [20]. Sentiment analysis is based on the polarity of the text in tweets to determine whether the text conveys positive, negative, or neutral sentiment. Sentiment can be formulated as:

$$Sentiment(T) = \{positive, negative, neutral\} \tag{4.2}$$

Where T is the *text* of the tweet. The following tweet example illustrates the difference between stance and sentiment:

> *It is sad that the vaccines are not available yet so we can get rid of COVID soon*

In this example, the tweet expresses a negative sentiment. However, the tweet takes a clear stance in favor of vaccination (the stance target).

Most stance detection studies focus on predicting and analyzing the stance conveyed in a tweet's text. The goal of these studies is to determine whether a given tweet's text favors, is against, or takes a neutral stance toward a target. Although this approach may be appropriate for different applications such as detecting fake news and veracity checking [73, 74, 75], it is not optimal for the application of measuring and analyzing public opinion with stance detection because users, and possibly social bots, can post

numerous tweets and change their opinions over time.

Inspired by collective sentiment research, which defines collective sentiment as the sum of individual sentiment [76, 77, 54, 78, 79], we aggregate the stance of individual tweets (tweet-based stance) and users (user-based stance). We emphasize that the aggregated stance is a way to apply a stance detection measure of public opinion over time. The following example illustrates how the difference between tweet- and user-based stances could arise.

A dataset has 100 tweets and only two users: A and B. User A tweeted 10 times in favor of a target G, and User B tweeted 90 tweets against the same target G on the same day. In this sample dataset, 90% of the tweets are in favor of the target, and 10% are against the target. However, 50% of the users favor the target, and 50% are against the target. The interest is in analyzing the stance of users instead of tweets for public opinion. Therefore, tweet- and user-based stance analyses can answer different questions for different applications. We argue that the tweet-based stance for measuring public opinion is equivalent to taking multiple responses from the same participant in the same survey or poll and that the user-based stance is sampling each person only once.

To better understand how stance detection can be used to measure public opinion from Twitter data, we examine the stance measured by aggregating the stance of *tweets* (tweet-based stance) and compare it with the stance measured by aggregating the stance of *users* (user-based stance). In this chapter, we focus on the following research question:

- **RQ)** How does the report of public opinion from Twitter data differ when based on the analysis of tweet-based stance vs user-based stance?

We address this research question by a statistical analysis of the differences between tweet- and user-based stances. Although most studies on stance detection focus on improving the performance of the stance detection task, we move beyond this to

understand how stance detection can be used to measure public opinion by introducing and investigating the difference between tweet-based and user-based stance analyses. We claim that user-based stance analysis is more aligned with the goal of using Twitter data as a way of measuring public opinion than tweet-based stance analysis.

The remainder of the chapter is organized as follows. In Section 2, we provide background and related work on opinion mining, stance detection, and stance analysis. In Section 3, we describe our data collection. In Section 4, we explain how we calculated the tweet-based and user-based stance for our analysis. In Section 5, we statistically compare tweet- based and user-based stance analyses and discuss the results. In Section 6, we outline our future work. We conclude in Section 7 and provide insight into the broader impact and ethical considerations of this research in this chapter.

4.2 Background and Related Work

This section provides background information related to opinion mining, stance, stance detection, and stance analysis.

4.2.1 Opinion Mining and Stance Detection

Sentiment analysis has long been associated with the mining of opinion of Twitter data. Several studies have used sentiment analysis to measure public opinion in different domains, such as public health [80], politics [81, 82, 83], and economy [84]. However, the success of sentiment analysis in measuring public opinion has varied. In one study, researchers investigated the analysis of collective sentiment as a predictor of the 2010 US Senate special election in Massachusetts [76]. Their analysis demonstrated that sentiment analysis performs similarly to random classification in predicting the results.

Stance detection, also known as stance prediction and classification, in Twitter data is the process of automatically determining the standpoint expressed in a tweet toward a target of interest, such as a person or an issue [32, 33]. The stance can be

in favor, against, or neutral. Unlike sentiment analysis, the goal of stance detection is to identify how a user positions themselves (explicitly or implicitly) with respect to a target, such as a presidential candidate. Applications of stance detection on social media platforms include the detection of false news, public health surveillance, and information retrieval [33]. For example, [39] used deep learning stance detection to detect fake news. However, the literature lacks studies exploring how stance detection can be used as a robust measure of public opinion in Twitter data.

Unlike sentiment analysis, limited studies have used stance detection to identify public opinion toward targets of interest. In one study, researchers used stance detection in Twitter data to measure public opinion toward the Job Creation Bill in Indonesia [48]. Another study by researchers investigated public opinion toward e-cigarettes in Indonesia [49].

However, these studies used the stance of tweets as a way to measure public opinion without considering that the user-tweet relationship is a one-to-many relationship. That is, a single user can tweet multiple times. One implication of a relationship of this nature is that a single user can introduce bias in the measurement of public opinion if they tweet with high frequency. Furthermore, the analyses in the previous research did not consider tracking the stance over time. Tracking the aggregated stance over time, with a temporal unit of analysis of a day, week, or month, is particularly important for tracking changes and dynamics of public opinion. This research bridges this gap by investigating how the daily tweet-based stance compares to the daily user-based stance over a period of 180 days.

4.2.2 Stance Analysis

Stance *analysis* can be either tweet-based or user-based. In one study, researchers followed a tweet-based stance analysis to analyze public opinion toward masks during the COVID-19 pandemic [85]. In another study, researchers investigated public opinion towards e-cigarettes in Indonesia [49]. These studies followed a tweet-based

stance analysis approach to determine public opinion toward specific topics without considering that the user-tweet relationship is a one-to-many relationship. That is, a single user can tweet multiple times. One implication of a relationship of this nature is that a single user can introduce bias in the measurement of public opinion if they tweet with high frequency. Furthermore, the analyses in the previous research did not consider tracking the stance over time. Tracking the aggregated stance over time, with a temporal unit of analysis of a day, week, or month, is particularly important for tracking changes and dynamics of public opinion. Similarly, there are a few studies that followed a user-based stance analysis approach to measure public opinion from Twitter data. In the study of [86], the researchers followed a user-based stance analysis as a measure of public opinion towards vaccination in Italy. In the work of [59], the authors followed a user-based stance analysis approach to gauge public opinion towards Brexit. It is evident that studies use tweet-based and user-based stance analyses to measure public opinion from Twitter data. However, it is unclear how the report of public opinion from Twitter data differs when based on the analysis of the tweet-based stance versus user-based stance. Figure 4.1 shows an example to illustrate the difference between the tweet-based and the user-based stance analyses.

The limitations of previous studies and the implications of using stance detection to measure public opinion motivated the research in this chapter. First, our goal is to determine the stance of the users at different points in time using only the content of their tweets. Second, we aim to compare the aggregated user-based stance with the aggregated tweet-based stance to examine whether they yield significantly different stance analyses.

4.3 Data

We used the data that we collected and described in Chapter 3. To recap, we collected tweets about vaccination using Twitter's streaming API between June 2 and November 28, 2021. During this period, some COVID-19 vaccines, such as the

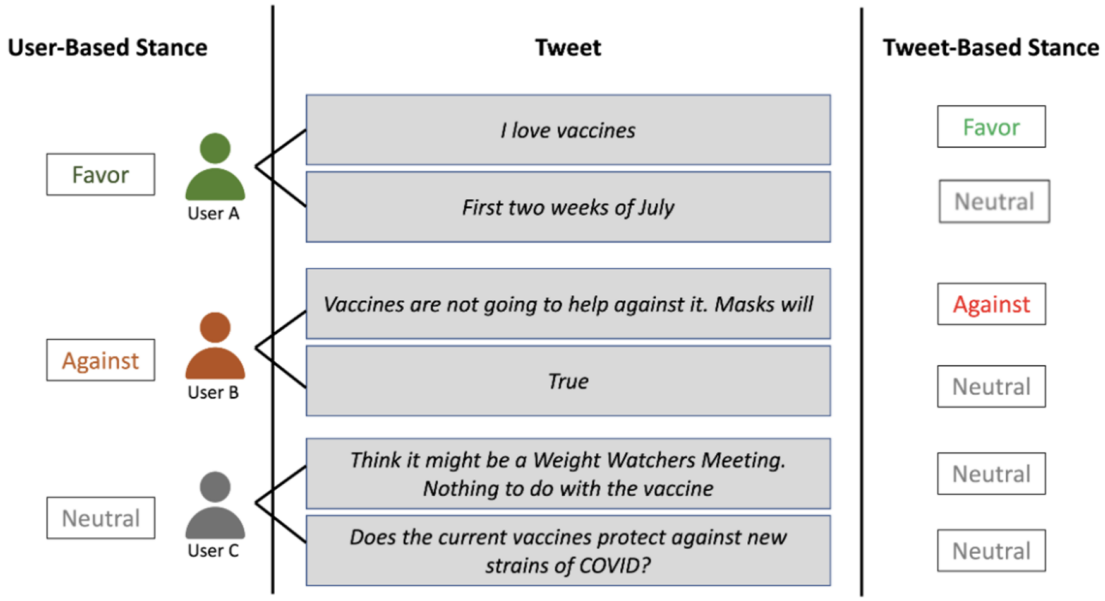

Figure 4.1: An example of the difference between tweet- and user-based stance in tweets

Pfizer and Moderna vaccines, were widely available to the population and people used Twitter to express their opinions about the vaccine. To ensure reliable data collection, we set up the data stream on Amazon Web Services. To filter the stream, we used general vaccine-related keywords and hashtags from the literature to examine pro- and anti-vaccination (ProVax and AntiVax) users on Twitter [71, 72]. The complete list of keywords and hashtags is presented in Table 3.1.

The data covered 180 days and excluded retweets and tweets with external links from the data collection. Tweets not in English were also excluded from the data collection stream. The final dataset had 24,806,152 tweets and 4,664,146 unique users. We show the percentage of users with tweeting frequency between 1 and 12 in Figure 4.2. We can see from the figure that half of the users tweeted only one tweet, and the other half tweeted more than one tweet.

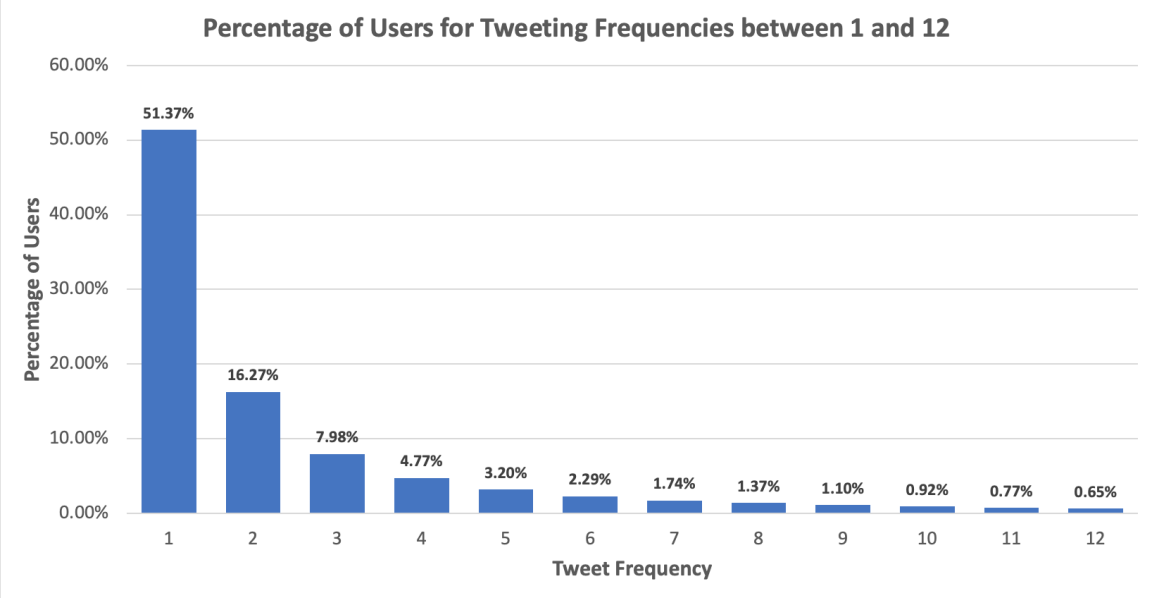

Figure 4.2: The percentage of users for tweeting frequencies between 1 and 12

4.4 Stance Classification and Aggregation

In this section, we describe our approach to detect the stance for the tweet-based and user-based stance analyses. Then, we show how we aggregated the stance for the temporal analysis.

4.4.1 Stance Classification

Stance classification on social media is a process of automatically labeling tweets or users with a stance label such as favor, against, or none toward the target. In our case, the stance target was vaccination. To answer our research questions, we relied on a tweet-based classifier to determine the stance. In a tweet-based stance classifier, the classifier takes single tweets as input and provides a stance label (favor/against/neutral) as output. We automatically classified the stance of each tweet towards vaccination using a CT-BERT++ state-of-the-art stance classifier [87] that is publicly available[1].

The CT-BERT++ classifier was pretrained on 5.7 million English tweets from 4

[1]https://github.com/sohampoddar26/covid-vax-stance

million distinct users over the duration between January 2018 and March 2021. This date range covers three different periods of time: 1) the pre-COVID period (January 2018 to December 2019), 2) the COVID period (2020), and 3) the COVID-Vax period (January to March 2021). The authors report an average macro F1 score of 0.775. This classifier was also validated using a different manually labeled Twitter dataset, resulting in a macro average F1 score of 0.83 for the ProVax and AntiVax stance classes [43]. The percentage of tweets with favor stance and against stance are shown in Figure 4.3.

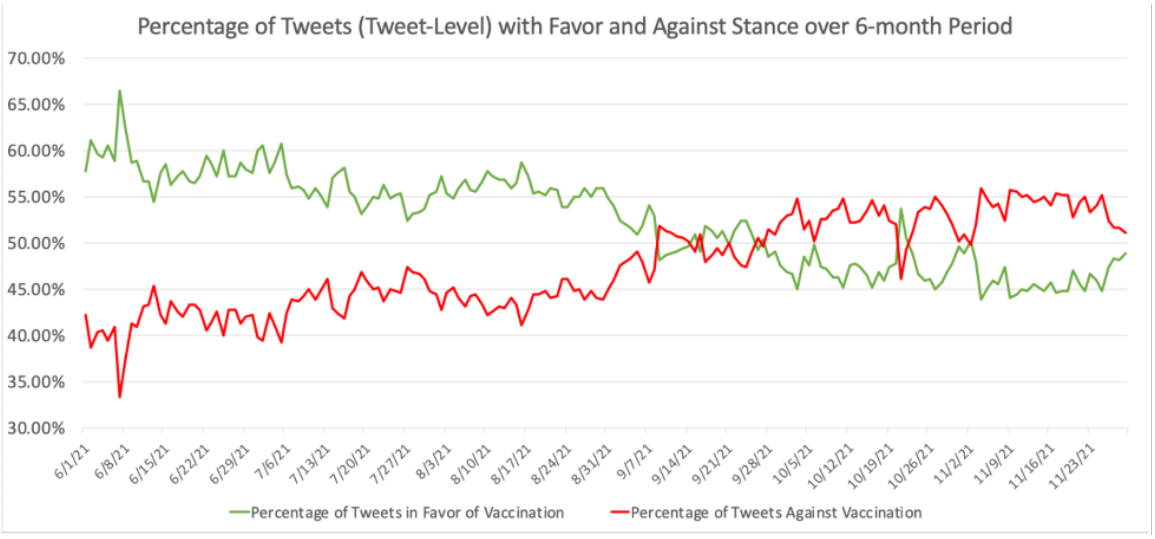

Figure 4.3: Stance over time for favor and against classes

We can see from the figure that in September 2021, the volume of tweets against vaccination overpassed the volume of tweets in favor of vaccination. By manually inspecting the tweets around the change, we observed that stance-taking at the time of the change was related to President Biden posting a series of tweets about new vaccine mandates for companies with specific criteria, as shown in his tweet in Figure 4.4.

Using the classifier, we classified all tweets in the dataset into three stance classes:

Figure 4.4: One of President's Biden tweets that triggered a change in the volumes of tweets in favor and against vaccination

ProVax, AntiVax, and Neutral. The percentage of each stance category in the data is shown in 4.5.

The following are examples of labeled tweets in the dataset.

> **ProVax Tweet:** *Anti vaccine rhetoric is dangerous. Hesitancy and refusal threaten public health. It's totally unacceptable. #VaccinesWork #stopantivaxviolence #COVID19 #COVID_19*

> **AntiVax Tweet:** *This #NWO confusion is counter-productive/suicidal to our health. We reject it #NWO We reject #Masks / #lockdown We reject #vaccines because #VaccinesWork - NOT WE REJECT HATE, FEAR - ALL ARE TO BE PURIFIED WITH TRUTH/LOVE OF HUMANITY - TRUE HUMANITY #GreatAwakeningWorldwide*

> **Neutral Tweet:** *Will you take COVID-19 vaccine once it becomes available? Take the poll!#COVID19 #VaccinesWork #CovidVaccine #CoronavirusVaccine #CovidUpdates*

To aggregate the stance, we used two units of aggregation: tweets (tweet-based stance aggregation) and users (user-based stance aggregation). We aggregated the data daily for each class of stances separately and calculated the total number of

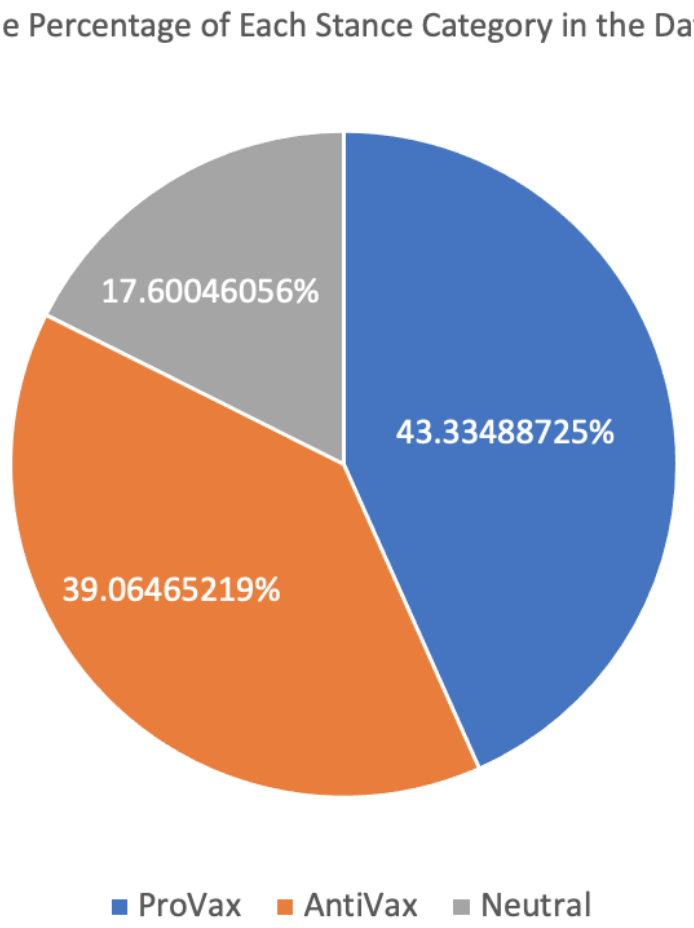

Figure 4.5: The percentage of each stance category in the data

data points in each class to the total number of tweets (for the tweet-based stance) or users (for the user-based stance). The following subsections explain the details of our stance aggregation.

4.4.2 Tweet-Based Stance Aggregation

We calculated the tweet-based stance for one day by calculating the ratio of tweets in each class (ProVax, AntiVax, and Neutral) as classified by the classifier to the total number of tweets that day. We measure the tweet-based stance ST for a stance class c on a specific day i as follows:

$$ST_{c,i} = \frac{t_{c,i}}{t_i} \tag{4.3}$$

where $c = \{ProVax, AntiVax, Neutral\}$, $t_{c,i}$ is the number of tweets that belong to class c for day i, and t_i is the total number of tweets for day i.

4.4.3 User-Based Stance Aggregation

We followed two steps to identify the user-based stance. First, we established the daily stance of the unique users from the stance of their tweets, where a user's stance is defined by the stance of the majority of their tweets for a specific day.

Next, we calculate the user-based stance SU for a stance class c on specific day i as follows:

$$SU_{c,i} = \frac{u_{c,i}}{u_i} \tag{4.4}$$

where $c = \{ProVax, AntiVax, Neutral\}$, $u_{c,i}$ is the number of users that belong to class c for day i, and u_i is the total number of users for day i. We assume that a user's stance will typically remain constant for a short period of time, such as 24 hours, and when there are specific days in which their stance changes, we can use the majority stance for that day.

The basis of our assumption is that public opinion at the individual level remains constant unless new information becomes available and triggers a change of opinion [88]. An example of such triggers for COVID-19 vaccination could be new official guidelines or news that are related side effects similar to those reported for the Johnson & Johnson vaccine and resulted in its suspension in the United States [89, 90]. Intuitively, we expect that longer periods of time will have more triggers of opinion change, especially for a fast-growing public issue such as COVID-19 vaccination.

To illustrate the aggregation method, consider the following example:

4.5 Tweet-Based vs User-Based Stance Analysis

Figure 4.6 illustrates the aggregated stance per day over 180 days and Figure 4.7 presents the frequency for each stance class for both tweets and users. Overall, we can

see that tweet-based and user-based stance aggregations follow similar trends, but for the purposes of this research, we want to examine if they are significantly different as the basis for the measurement of public opinion. We examine the differences in the distribution of tweet- and user-based stances using statistical and visual methods. The properties of the distributions are crucial in selecting the appropriate statistical test to compare tweet-based and user-based stance analyses.

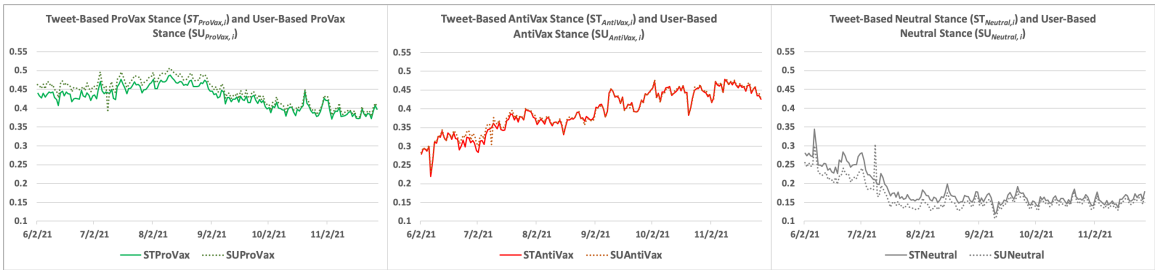

Figure 4.6: Tweet-based vs user-based ProVax (left), AntiVax (middle), and Neutral (right) stances per day between June 2, 2021 and November 28, 2021

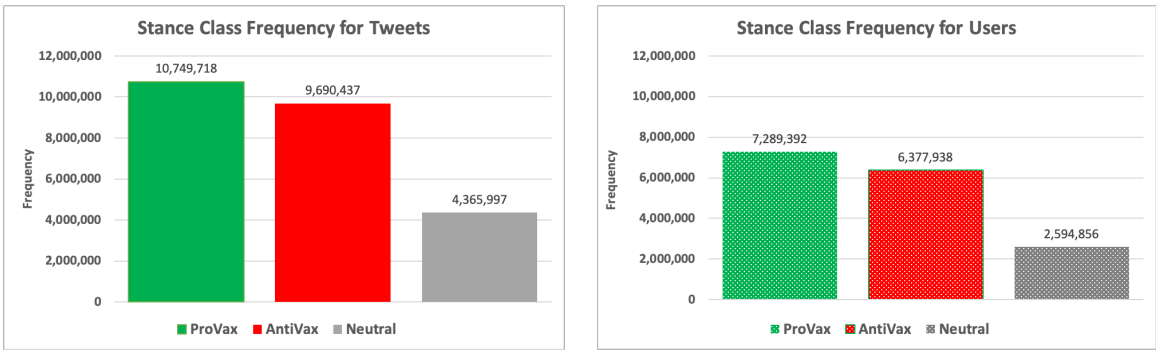

Figure 4.7: Stance class frequency for both the tweets (left) and the users (right)

4.5.1 Checking for Data Normality

Data normality is a property of a variable that follows a normal distribution. Many statistical tests such as the t-test assume normality. To check the normality of the data, we use the Kolmogorov-Smirnov (K-S) normality test [91]. The K-S test is a one-sample nonparametric test that provides a measurement of the divergence of a sample's distribution from the normal distribution with the following null hypothesis:

Table 4.1: Summary of K-S test statistics for tweet-based stance

	ProVax	AntiVax	Neutral
Count	180	180	180
Mean	0.42784	0.39058	0.18157
Median	0.429438	0.390867	0.164495
SD	0.029947	0.055406	0.041865
Skewness	-0.100295	-0.36942	1.419419
Kurtosis	-0.998642	-0.673718	1.159148
K-S test statistic	0.0665	0.10706	0.25125
p-value	0.38681	0.02989	<0.00001
Distribution	Normal	Non-Normal	Non-Normal

Table 4.2: Summary of K-S test statistics for user-based stance

	ProVax	AntiVax	Neutral
Count	180	180	180
Mean	0.44243	0.39353	0.16404
Median	0.444912	0.392205	0.150685
SD	0.03488	0.053119	0.036098
Skewness	-0.207364	-0.317241	1.542597
Kurtosis	-1.140491	-0.624512	1.994866
K-S test statistic	0.08496	0.10571	0.21106
p-value	0.14042	0.03319	<0.00001
Distribution	Normal	Non-Normal	Non-Normal

Null hypothesis (1) H_0**:** The data are not significantly different from a normal population.

Where a p-value>0.05 indicates that the data are normal. We present a distribution summary and the K-S test statistics for tweet-based and user-based stances in Tables 4.1 and 4.2, respectively. The results of the K-S normality test indicated that the data follow a normal distribution for the tweet-based and user-based ProVax stance class because the p-value is greater than 0.05. However, the AntiVax and Neutral stance classes do not follow a normal distribution for tweet- and user-based stance classes because the p-value is less than 0.05.

Statistical tests of normality, such as the K-S test, are useful for examining the distribution of the data. However, they have limitations. One limitation of the K-S

test is that it tends to be more sensitive near the center of the distribution than at the tails. Due to this limitation, other methods to assess normality should be considered. In this chapter, we complement our statistical tests of normality with two visual methods: distribution plots and quantile-quantile (q-q) plots.

Distribution plots are widely used to assess the normality of the data. The distribution plots show the frequency at which each value appears in the data. The normal distribution can be recognized because of its bell shape on a distribution plot. From Figure 4.8, we can see that the data do not form a bell shape. In contrast, both the ProVax and AntiVax stance classes appear to have a bimodal distribution, which is characterized by two peaks. The neutral stance skewness also indicates that the data do not follow a normal distribution.

Quantile-quantile plots are useful for comparing the distribution of a variable with a selected distribution, such as the normal distribution. For data drawn from a normal distribution, the data in the q-q plot are expected to form a straight line along the 45-degree line. Visual inspection of the quantile-quantile plots (as presented in Figure 4.9) indicated that the data clearly diverge from the 45-degree line for the neutral stance for both the tweet-based and the user-based stance analyses. In addition, the data were not close to forming a straight line as would be expected in a normal distribution. Because this finding aligns with the results from the K-S test, we treated the distribution of the neutral stance class as a nonnormal distribution. In contrast to the Neutral class, it was unclear whether the ProVax and AntiVax classes follow a normal distribution. Although the values for both classes were clustered along the 45-degree line, they do not form a straight line.

4.5.2 Testing for Significant Difference

Based on our test of normality, we cannot conclude that the data are normal for any of the ProVax, AntiVax, and Neutral stance classes. Therefore, we elected to use a nonparametric test to assess the significant difference between tweet-based and

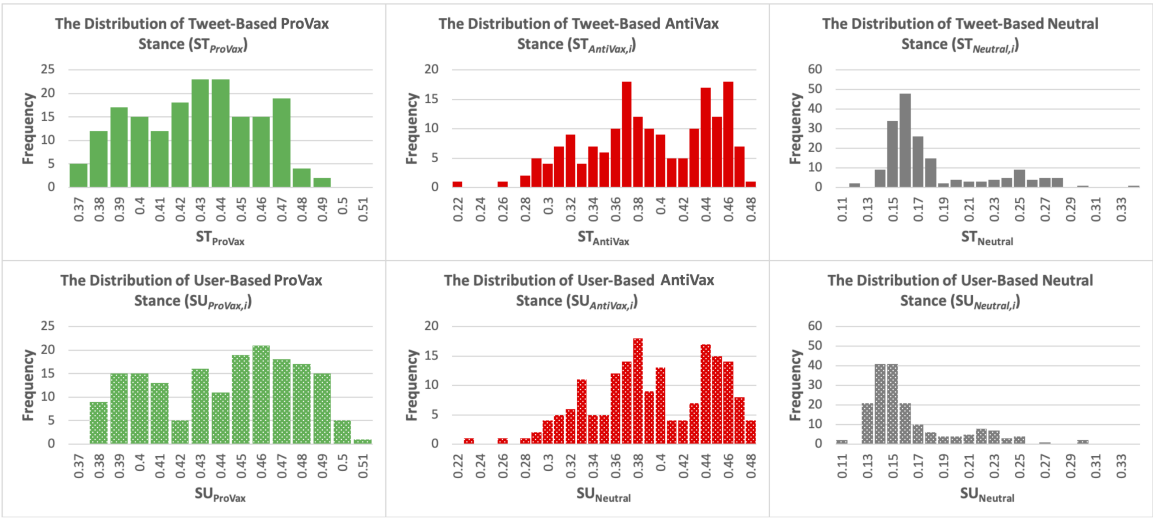

Figure 4.8: The distribution of tweet-based stance (top) and user-based stance (bottom) for ProVax (left), AntiVax(middle), and Neutral (right) stance classes

user-based stance analyses. Non-praprametric tests do not assume data normality. A well-known nonparametric test to compare outcomes between two independent groups is the Wilcoxon Signed-Rank Test. The Wilcoxon Signed-Rank Test is a paired test that is used to test whether two samples are likely to derive from the same population [92]. We used the Wilcoxon Signed-Rank Test with the following null hypothesis:

Null Hypothesis (2) H_0**:** The two populations are equal.

We used the test as a two-sided test. That is, our main interest was to identify whether the two populations were not equal without specifying directionality. The test results from the evaluation of the difference between tweet-based and user-based stance analyses for the AntiVax and Neutral stance classes are shown in Table 4.3. From the results, we can see that the p-values for the ProVax, AntiVax, and Neutral stance classes are less than 0.5, indicating a significant difference. Therefore, we reject Null Hypothesis (2) that the two populations are equal. This is statistically significant evidence that the tweet-based stance analysis is not equal to the user-based stance analysis.

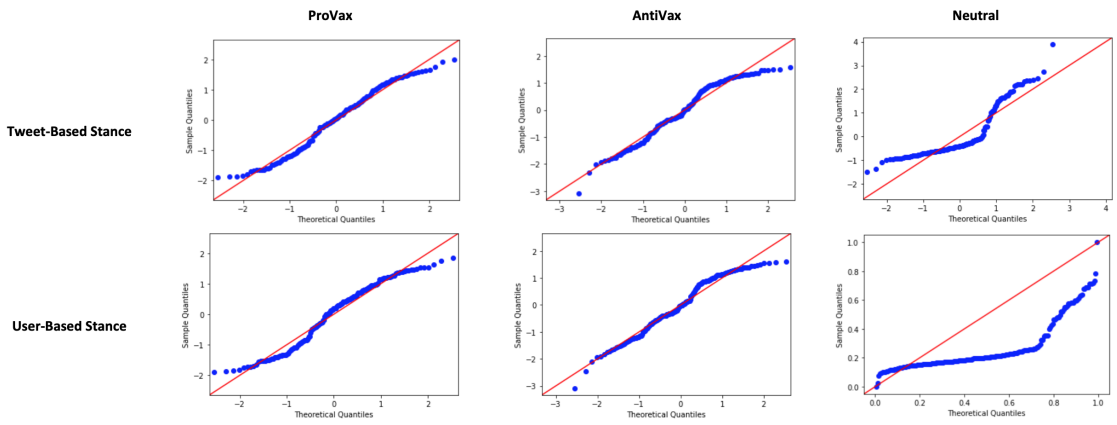

Figure 4.9: Quantile-quantile plots for visually comparing the distributions of ProVax (left), AntiVax (middle), and Neutral stance classes to normal distribution

4.5.3 Effect Size Analysis

While assessing statistical significance shows that there is a difference, practical significance is useful to assess whether the difference is large enough to be meaningful. Effect size is a quantitative measure of the magnitude of the effect [93]. There are different indices that can be used to assess the effect size. In this research, we use Cohen's d and Pearson's correlation r to examine the effect size. Cohen's d is an appropriate effect size measure for the comparison between two means. Cohen's d is calculated as follows:

$$d = \frac{M_1 - M_2}{SD_{pooled}} \tag{4.5}$$

Where $M1$ is the first mean, $M2$ is the second mean, and SD_{Polled} is the average standard deviation. Cohen's d values can be interpreted as:

- $d = 0.2$ indicates a small effect
- $d = 0.5$ indicates a medium effect
- $d = 0.8$ indicates a large effect

The values of Pearson Correlation r range between -1, indicating a strong negative

Table 4.3: The Wilcoxon Signed-Rank Test Statistics for Testing the Difference Between Tweet-Based and User-Based Stance

	ProVax	**AntiVax**	**Neutral**
W-value	180	3547	180
Mean Difference	-0.03	0.09	-0.11
Sum of Positive Ranks	180	3547	180
Sum of Negative Ranks	16110	12743	16110
Z-Value	-11.3779	-6.5682	-11.3779
Mean (W)	8145	8145	8145
Standard Deviation (W)	700.04	700.04	700.04
Sample Size N	180	180	180
p-value	$< .00001$	< 0.00001	$< .00001$
Result	Significant	Significant	Significant

correlation, and 1, indicating a strong positive correlation. We present the results of the evaluation of the effect size in Table 4.4. We can see from Table 4.4 that cohen's d and the Pearson correlation r vary per stance class. They concluded that the effect size of ProVax and Neutral is medium with d = 0.44 and 0.448. However, the effect size of the AntiVax is small. This finding aligns with the visualization of the difference in Figure 4.6, where we see a clear difference between the tweet-based and user-based stance analyses for the ProVax and Neutral stance classes, and the difference is small for the AntiVax stance class.

Table 4.4: Cohen's d and Pearson's correlation r for ProVax, AntiVax, and Neutral stance classes

	Cohen's d	**Pearson's r**
ProVax	0.440	0.219
AntiVax	0.054	0.027
Neutral	0.448	0.219

4.6 Summary

In this research, we distinguished between tweet-based and user-based stance to measure public opinion from Twitter data. Our analysis of comparing tweet- and

user-based stance analyses showed that the two analyses produce statistically different results. That is, if public opinion is measured from the aggregation of tweets, the public opinion will be different from the public opinion measured from the aggregation of users. Although we found that the difference varied from one stance class to another, we conclude that the difference is statistically different. This work provides a methodology for measuring the user-based stance and analyzing whether it is significantly different from the tweet-based stance for gauging public opinion.

4.7 Broader Impact and Ethical Considerations

This research uses stance detection on social media data as a measure of public opinion. One positive outcome of this research is that it uses widely available and accessible social media data to analyze public opinion as an alternative to traditional expensive and time-consuming methods, such as interviews. With surveys and interviews, researchers cannot gauge public opinion about an event that occurred in the past or an unfolding event in the meantime. However, social media data can be collected from the past to analyze public opinion and how public opinion evolves over time.

Although social media data are a useful instrument for such an application, we recognize that there are ethical considerations. Unlike traditional forms of surveys and polls, the acquisition of Twitter data lacks informed consent from users. As a result, Twitter users were unaware that their posts can be used to gain insight into public opinion on topics of interest, such as vaccination. To mitigate any negative impact, the results of the analysis do not include information that could lead to the identification of Twitter users. Therefore, the risk of negative consequences is very low.

CHAPTER 5: USER-BASED STANCE ANALYSIS FOR MITIGATING THE IMPACT OF SOCIAL BOTS ON MEASURING PUBLIC OPINION WITH STANCE DETECTION IN TWITTER

5.1 The Influence of Social Bots on Public Opinion

Social media platforms provide a medium for users to freely express their opinions about various issues [33, 32]. Although sentiment analysis, or opinion mining, has long been used to analyze public opinion, recent studies have demonstrated a misalignment between text polarity and the user's perspective on the target of interest [94, 95, 96]. On the other hand, stance detection is argued to serve as a better measure of public opinion. Although Twitter data are a valuable data resource for analyzing public opinion, there are implications of using Twitter, and one implication is the presence of social bots [97].

To understand the impact of social bots on the measured public opinion and how such impact can be mitigated, we analyze the stance by using two analysis approaches: tweet-based and user-based stance analyses. In a tweet-based approach, the stances of favor, against, and neutral *tweets* are aggregated and visualized separately at the daily level. Meanwhile, a user-based stance analysis approach places *user* at the center of the analysis by inferring their stance of the user from their tweets, as users are the unit of analysis for public opinion. Therefore, the stances of favor, against, and neutral users are aggregated and visualized. In this chapter, our aim is to answer the following research question:

- **RQ)** How does a user-based stance analysis mitigate the impact of social bots on measured public opinion in Twitter?

To answer the research question, we collect English tweets and bot scores associated with accounts contributing to the discussion around vaccination on Twitter and analyze the data using these two approaches. This study provides insight for researchers to minimize the impact of bots on their measurement of public opinion and improve the robustness of their measurement of public opinion using stance detection.

5.2 Background and Related Work

The following subsections provide background on social bots and stance detection.

5.2.1 Social Bots and Public Opinion

Social media research uses the term social bots to describe automated programs that exhibit human-like behavior and generate and spread content on social media [97]. Social bots can be benign or malicious. Benign bots (also called good, helpful, and useful) automatically generate helpful information, such as news and weather reports. Meanwhile, malicious bots are widely used to spread false information and influence public opinion. Social bots can influence public opinion by tweeting and retweeting content on Twitter.

Recent research studies suggest that bots contribute to the discussion of sensitive topics, such as political and vaccination debates [98, 99]. In 2016, bots played a substantial role in manipulating public opinion in the US presidential elections [100, 101]. In particular, a study [102] revealed that bot accounts tweet with a higher frequency than non-bot accounts. To our knowledge, the literature lacks studies investigating the impact of bots on tweet-based and user-based stance analysis as measures of public opinion.

In general, a bot (short for robot) is an automated program that interacts with humans. In the social media context, social bots are designed to mimic the behavior of human accounts on platforms such as Twitter (i.e., Twitter bots). The design of such bots has become more sophisticated over time. As mentioned, these bots

can influence public opinion by tweeting and retweeting content, with a recent study revealing that approximately 14% of the accounts contributing to the COVID-19 discussion on Twitter were likely bots [103]. This is emblematic of the way that some social bots can be dangerous to the public. For example, [104] analyzed tweets related to COVID-19 vaccination and found that bots were responsible for posting 11% of the tweets in their corpus. In the study of [105], the authors reported that bots were responsible for spreading low-credibility content related to COVID-19 vaccination.

5.2.2 Stance Detection

Stance detection, also known as stance prediction and classification, in Twitter data is the task of automatically determining the standpoint expressed in a tweet towards a target of interest, such as a person, topic, organization, issue or claim [33, 32]. Stance detection is different from sentiment analysis, another well-known NLP task, where the latter focuses on determining the polarity of the text in a tweet [20]. A fundamental difference between the two tasks is that sentiment analysis is designed to determine the polarity of the tweet (positive, negative, or neutral). On the other hand, stance analysis intends to determine the users' attitude toward the target based on their tweets (favor, against, neutral).

Most stance detection studies are concerned with improving the performance of stance detection algorithms [106, 107, 108, 12, 109]. However, few studies have looked beyond the algorithm and used stance detection for the application of measuring public opinion. A study followed a tweet-based stance analysis to analyze public opinion toward masks during the COVID-19 pandemic [85]. In another study [86], the researchers followed a user-based stance analysis as a measure of public opinion toward vaccination in Italy. In the work of [59], the authors followed a user-based stance analysis approach to gauge public opinion toward Brexit. However, previous studies have not considered the impact of social bots on tweet-based and user-based stance analyses.

5.3 Data Collection and Classification

In recent years, several tools have been developed to enable researchers to detect the activity of social bots. In this research, we used Botometer, a well-known bot detection tool for the research community for Twitter data [110]. Rather than assigning definitive labels (bot or human) to Twitter accounts, Botometer assigns numerical values to accounts to indicate the likelihood that the account is automated (i.e., the likelihood of the account being a bot). Botometer assigns scores between 0 and 1, where 0 indicates that the account is unlikely to be automated and 1 indicates that the account is likely to be automated. Botometer also provides the same scores on a scale between 0 and 5 for the graphical user interface, as Figures 5.2 and 5.2 show.

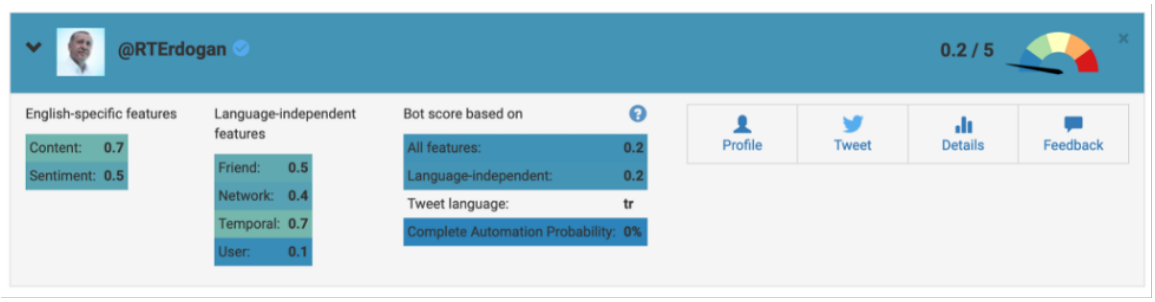

Figure 5.1: Botometer's graphical user interface indicating that an account is unlikely to be a bot

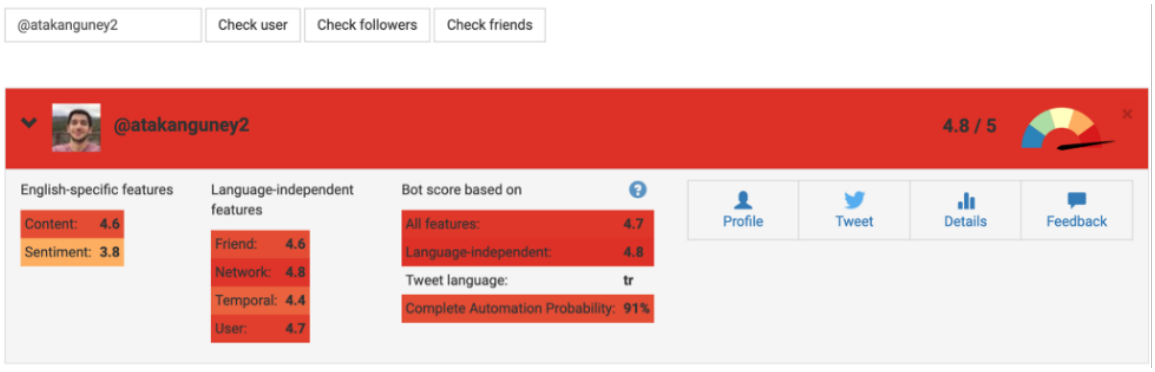

Figure 5.2: Botometer's graphical user interface indicating that an account is most likely to be a bot

The botometer employs a variety of features. The English score takes into consid-

eration the linguistic features of user tweets, and the Universal score considers user features only (for example, how often an account posts). Although we have presented studies demonstrating that bots are prevalent on Twitter and discussed the tools available for researchers to detect bots, it is unclear how the presence of bots affects the quality of measuring public opinion with stance detection. This chapter presents a novel approach that incorporates bot detection into stance detection to measure public opinion. Our approach allows researchers to separately analyze the stances of bot accounts and non-bot accounts. Note that this chapter uses the terms bots and social bots interchangeably to refer to bots operating on social media.

To identify bot accounts in our data, we used Botometer V4 to assess whether an account exhibits automated behavior [110]. Because our data comprises only English-language tweets, we have used Botometer's English scores, ignoring the universal scores.

However, this research does not use the numerical bot scores provided by Botometer. Instead, we have assigned each account a binary label (bot or non-bot) based on the Botometer English score. This required selecting a threshold. For this purpose, we reviewed Botometer's instructions and graphical interface and used the following threshold:

- 0.00 <= score <= 0.39: Likely not bot (blue and green meter colors)
- 0.40 <= score <= 0.60: Unclear (yellow meter color)
- 0.61 <= score <= 1.00: Likely bot (orange and red meter color)

To minimize the impact of bot misclassification on the analysis, we considered accounts with bot scores below 0.40 to be non-bot accounts and accounts with bot scores above 0.60 to be bot accounts, discarding ambiguous accounts with bot scores between 0.40 and 0.60).

To address this chapter's research question, we used Botometer on the Twitter data we collected and described in Chapter 2. To recap, we collected tweets related to the vaccination discussion using Twitter's streaming API between June 2 and November 28, 2021. During this period, some COVID-19 vaccines, such as the Pfizer and Moderna vaccines, were widely available to the population and people used Twitter to express their opinions about the vaccine. We elected to use the streaming API as opposed to the historical API to be able to collect account information for accounts that were subsequently deleted or suspended. To ensure reliable data collection, we established the data stream on Amazon Web Services (AWS). To filter the stream, we used general vaccine-related keywords and hashtags from the literature to examine pro- and anti-vaccination (ProVax and AntiVax) users on Twitter. The complete list of keywords and hashtags appears in Table 3.1. The final dataset contains 15,813,362 tweets (without retweets and quotes) and 3,286,474 unique users.

Our exploratory analysis revealed that bot accounts do not survive suspension for a long period of time, especially if they violate Twitter's terms of service. This aligns with the findings of [102]. To ensure that we collected bot scores before they became unavailable due to deletion or suspension of accounts, we retrieved bot scores as tweets were streamed. As soon as we streamed the data, we used our AWS architecture to fully automate bot-score collection and processing and started collecting bot scores on August 7, 2021. Once a Twitter data file was complete (i.e., reached 100,000 tweets), a program would the file to maintain a list of its unique users. Then, the program would feed the list of unique users to Botometer to check the likelihood of the account being a bot. This method revealed that only 1% of all accounts had been deleted or suspended. Therefore, our final dataset included 3,245,504 unique users.

At the time of this chapter's study, Botometer offered three subscription plans: Basic, Pro, and Ultra. All three subscription plans had a limit on the number of accounts that could be checked per day. The most advanced plan, Ultra ($50/month),

allowed subscribers to check up to 17,280 accounts per day. After reaching the maximum limit in the Ultra plan, Botometer charges $0.001 per additional account check. In addition, Botometer's API features an average latency of 1,530 ms. Due to the large number of unique accounts in the data and this high latency, we opted to use multiple subscriptions simultaneously to ensure that we could collect the bot scores for all unique accounts at the same rate as the Twitter stream. Thus, we had a total of 11 active Ultra subscriptions dedicated to checking bot scores while the data were streamed from Twitter's API to AWS, enabling us to reliably check up to 190,080 accounts per day. This limit was sufficient to check the unique accounts that were being streamed.

Stance detection involved using stance labels specific to vaccination: ProVax, AntiVax, and Neutral. We automatically classified the stance of each tweet toward vaccination using a publicly available state-of-the-art stance classifier [111] [1]. The CT-BERT++ classifier was trained on tweets before COVID-19 and during COVID-19 (between 2019 and 2021), and the authors reported an average macro F1 score of 0.775. This classifier was also validated using a different manually labeled Twitter dataset, producing an average macro F1 score of 0.83 for the ProVax and AntiVax stance classes [43]. Using the classifier, we classified all tweets in the dataset into three stance classes: ProVax (43%), AntiVax (39%), and Neutral (18%).

5.4 Analysis and Discussion

Next, we analyzed tweet-based and user-based stance analyses as measures of public opinion with stance detection. Because public opinion is more associated with users than tweets, our goal is to investigate whether a user-based approach can mitigate the impact of social bots on measured public opinion.

First, we examined the presence of bots in our data. We note that on average 15% of the accounts that tweeted on a particular day were bot accounts. We also note

[1] https://github.com/sohampoddar26/covid-vax-stance

Table 5.1: Examples of tweets with ProVax, AntiVax, and Neutral stances posted by bot and non-bot accounts

Tweet	**Account Type**	**Stance**
Wouldn't it be easier, cheaper, and healthier for everyone to get the vaccine so we don't need these ambulances? We could wear masks until that happens.	Non-bot	ProVax
No. The vaccine is a magnetic device to track us	Non-bot	AntiVax
I thought she was comparing vaccines.	Non-bot	Neutral
The evidence that the vaccine is SAFE all the studies that make up the FULL CLINICAL TRIALS - which any vaccine or drug has before it goes to market	Bot	ProVax
They're not "vaccine passports," they're movement licenses. It's not a vaccine, it's experimental gene therapy. "Lockdown" is at best completely pointless universal medical isolation and at worst ubiquitous public incarceration. Call things what they are, not their euphemisms	Bot	AntiVax
BREAKING: San Francisco to require proof of COVID vaccine to enter restaurants, bars, gyms, etc., becoming first major U.S. city to do so	Bot	Neutral

that bot accounts were present in all ProVax, AntiVax, and Neutral stance classes. Table 5.1 shows examples of tweets posted by bot and non-bot accounts.

Next, we calculated the tweet-based stance for one day by calculating the ratio of tweets in each class (ProVax, AntiVax, and Neutral) as classified by the classifier to the total number of tweets that day. We measure the tweet-based stance ST for a stance class c on a specific day i as follows:

$$ST_{c,i} = \frac{t_{c,i}}{t_i} \tag{5.1}$$

where $c = \{ProVax, AntiVax, Neutral\}$, $t_{c,i}$ is the number of tweets that belong to class c for day i, and t_i is the total number of tweets for day i.

We followed two steps to identify the user-based stance. First, we established the daily stance of the unique users from the stance of their tweets, where a user's stance

is defined by the stance of the majority of their tweets for a specific day.

Next, we calculate the user-based stance SU for a stance class c on specific day i as follows:

$$SU_{c,i} = \frac{u_{c,i}}{u_i} \tag{5.2}$$

where $c = \{ProVax, AntiVax, Neutral\}$, $u_{c,i}$ is the number of users that belong to class c for day i, and u_i is the total number of users for day i. We assume that a user's stance will typically remain constant for a short period of time, such as 24 hours, and when there are specific days in which their stance changes, we can use the majority stance for that day. Tweet-based and user-based stance analyses with and without bots are shown in Figure 5.3. The blue dotted lines represent the stance without bots, and the orange solid lines represent the stance without bots.

From Figure 5.3 (left), we observe that there is a clearly visible difference between the tweet-based stance analysis with bots and the tweet-based stance analysis without bots for the ProVax, AntiVax and Neutral stance classes. This is evidence that tweet-based stance analysis is sensitive to the presence of bots. In contrast, Figure 5.3 (right) shows that the user-based stance analysis with bots and the user-based stance analysis without bots are nearly identical. By definition, bots are designed to automate the spread of content. Therefore, bots are expected to tweet with a high frequency. Because the tweet-based stance analysis considers *individual tweets*, regardless of whether they were posted by the same account, frequency has an impact on measuring public opinion. On the other hand, user-based stance analysis is more robust to the presence of bots because *individual users* are considered in measuring public opinion regardless of the tweeting frequency of users.

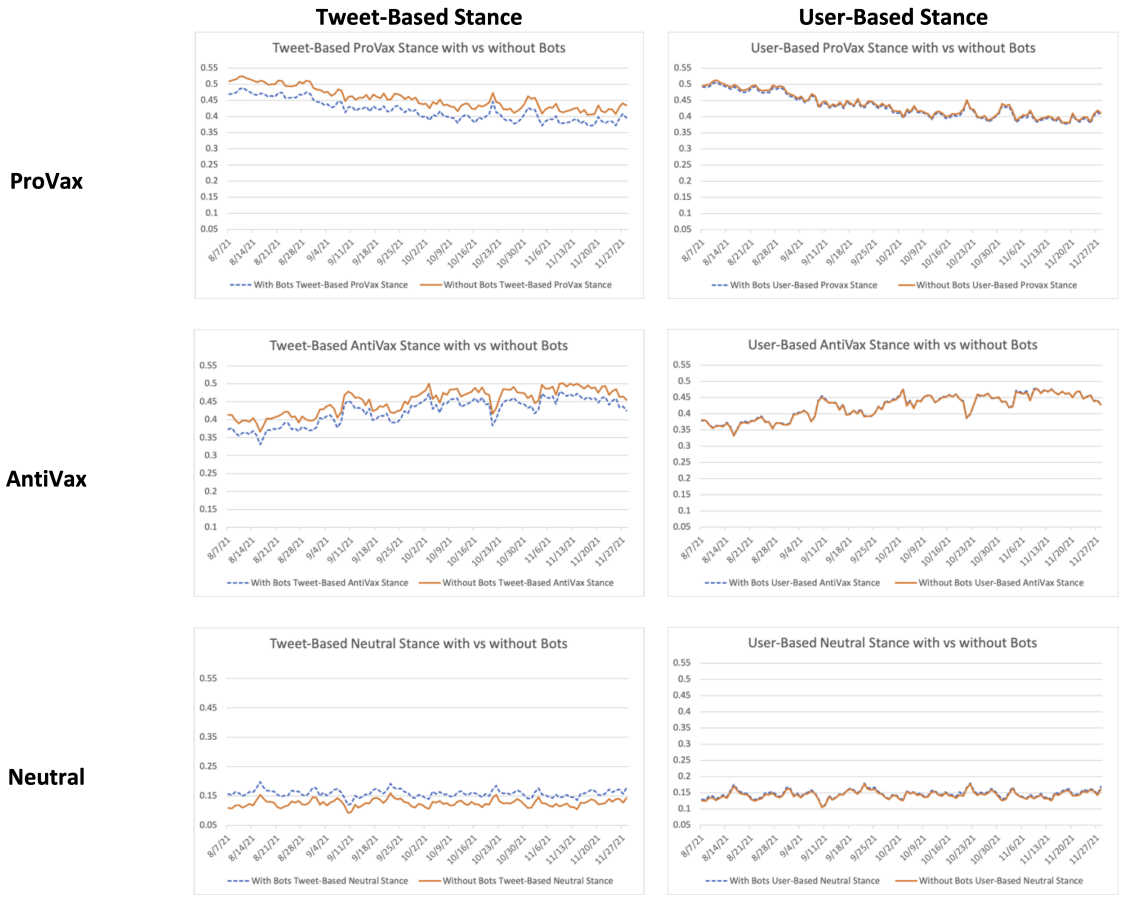

Figure 5.3: Comparison between tweet-based and user-based tweet analysis before and after removing bots

5.5 Summary

In this chapter, we compared two approaches for measuring public opinion with stance detection from Twitter data: tweet-based and user-based stance analyses. Although we found that tweet-based stance is sensitive to bots, there was a minimal impact of bots on measuring public opinion with user-based stance analysis. The results of this study provide information and considerations for researchers who intend to use stance detection for the application of measuring public opinion from Twitter data. Although we showed that there was a difference between tweet-based and user-based stance analyses in mitigating bots' impact, we emphasize that bots can

have different roles in different topics and discussions. Therefore, there is a need to investigate how a user-based approach mitigates the impact of bots for other stance targets and datasets.

CHAPTER 6: STANCEDASH - A DASHBOARD TO STUDY AND ANALYZE PUBLIC OPINION WITH STANCE DETECTION IN TWITTER DATA

6.1 The Need for Meaningful Stance Data Presentation

Social media platforms, such as Twitter, have gained increasing popularity in recent years. Social media provides a medium for people to freely express their opinions on various topics and issues without social burden. Therefore, Twitter data is viewed as a valuable source for understanding different phenomena such as public opinion. Natural Language Processing tasks such as emotion detection and sentiment analysis have long been associated with understanding how users feel on these platforms about a certain entity. However, these two approaches are arguably not aligned with the concept of public opinion. On the other hand, stance detection in Twitter data is defined as the task of automatically identifying the standpoint if a user toward a target of interest. Although, by definition, stance detection is the closest task to measuring public opinion in Twitter data, only a few studies investigated how data from stance detection can be visualized to allow studying and monitoring public opinion from Twitter data. An explanation for this limitation is perhaps the novelty of stance detection.

Due to the large volume of Twitter data, one of the significant challenges is representing the data to end users in a meaningful way to enable them to monitor, analyze, and understand public opinion. Building on our research on user-based stance analysis, the aim of this study is to present and evaluate a dashboard that assists end users analyze public opinion from stance detection in Twitter data and identify usability issues. Our main research question is:

- **RQ)** How does visualizing a user-based stance in a dashboard assist end users monitor, analyze, and understand public opinion in Twitter?

To answer this research question, we designed StanceDash, a web-based dashboard that presents and visualizes user-based stance in Twitter data. We conducted a study with 13 participants in a controlled environment to evaluate the dashboard and assess its usability using a mix of quantitative and qualitative analyses.

6.2 Relevant Research on Stance Visualization

With the exponential growth of social media, the need for visual text analytics has grown to interpret data to help understand large data. Visual text analytics have allowed one to extract information such as linguistic patterns.

In recent years, stance visualization has attracted many researchers in the field of text visualization in social media. The study of [41] developed one of the first interactive stance visualizations for SemEval 2016 Task 6. The tweet-based stance interactive tool provides a utility to visualize the percentages and counts for each stance category (favor/against/neutral) for 6 stance targets: Hilary Clinton, the Feminist Movement, Legalization of Abortion, Atheism, Donald Trump, and Climate Change is a Real Concern. The goal of the tool is to describe the data rather than helping end users analyze public opinion. Therefore, it has very limited functionality.

In another study [112], researchers developed StanceXplore, an interactive tool to inspect stance in Twitter data. StanceXplore is a tweet-based visualization tool with integrated temporal and geographical functionalities to understand the stance of tweets over time in a time view. StanceXplore has 5 views to support stance exploration: stance view, hashtag view, map view, timeline view, and tweets view. The stance view summarizes the number of tweets in each stance category. The hashtag view shows the most frequent hashtags in the data and allows users to filter the data by selecting hashtags. The map view shows a heat map, where countries with more tweets are distinguished by a dark red color. The timeline view shows the

number of tweets for each stance category daily. The tweet view provides a utility to examine the content of tweets. An modified version of StanceXplore is StanceVis Prime [113]. StanceVis Prime is also a tweet-based interactive stance and sentiment visualization tool that was developed in 2020. It supports data visualization from both Twitter and Reddit. Similar to StanceXplore, StanceVis Prime enables users to perform temporal analysis.

The previous stance visualization studies suffer from some limitations. While numerous research have shown that social bots contribute to the discussion of topics such as vaccinations, current visualization studies do not provide a utility for end users to analyze the stance of bots separately from the stance of non-bots. Nonetheless, the visualization tools have not been evaluated by potential end users to assess their usefulness and usability. In this research, we design and evaluate a web-based dashboard to overcome these limitations. To the best of our knowledge, this is the first study to design and evaluate a dashboard that assists end users analyze and monitor public opinion with stance detection in Twitter data.

6.3 Data

In this study, we used the data we described in Chapter 3. To recap, we collected tweets related to vaccination using Twitter's streaming API between June 2, 2021 and November 28, 2021. During this period, some COVID-19 vaccines, such as the Pfizer and Moderna vaccines, were widely available to the population and people used Twitter to express their opinions about the vaccine. To ensure reliable data collection, we set up the data stream on Amazon Web Services (AWS). To filter the stream, we used general vaccine-related keywords and hashtags from the literature to examine pro- and anti-vaccination (ProVax and AntiVax) users on Twitter. The complete list of keywords and hashtags is presented in Table 3.1.

The data cover 180 days and exclude retweets and tweets with external links from the data collection. Tweets not in English were also excluded from the data collection

stream. The final dataset had 24,806,152 tweets and 16,262,186 unique users. 51% of the users tweeted only one tweet and 49% of the users tweeted between 2 and 223,582 tweets.

6.4 StanceDash Design

To design StanceDash, we relied on design and evaluation studies from the literature on dashboard design. Our observation is that there are no specific design guidelines that are agreed upon for dashboard design. However, different studies have shown that an effective dashboard should be consistent and should not overwhelm the user with too much data [114, 115, 116]. Therefore, we considered these aspects as our design principles for StanceDash. The following two subsections describe the design goals, functionality, and components of StanceDash.

6.4.1 Design Goals and Functionality

Our design of the dashboard was guided by the system functionality with the end goal of providing end users with a utility that can assist them analyze and monitor public opinion toward certain targets of interest, such as vaccination. In particular, we had two predefined main design goals:

- **G1:** End users must be able to visualize and interpret public opinion and changes in public opinion as expressed by the stance of Twitter users (bots and non-bots) toward a target of interest for different times and locations
- **G2:** End users must be able to understand the language and topics of discussion that are associated with public opinion toward a target of interest for different Twitter users' types (bots and non-bots), times, and locations

These two design goals provide the essential functionality that StanceDash should provide. For example, to understand public opinion changes, some of the dashboard's visualizations should allow end users to conduct a temporal analysis of the users'

stance. In addition, the dashboard should provide data filters according to the types and locations of users.

6.4.2 StanceDash Components

The graphical user interface of StanceDash was adopted from the design of SB Admin 2, a free dashboard template that is available online. online[1]. The dashboard, as shown in Figure 6.1, consists of six components: navigation, data selection, summative data, stance data, bot data, and qualitative data components. Next, we describe each component in detail.

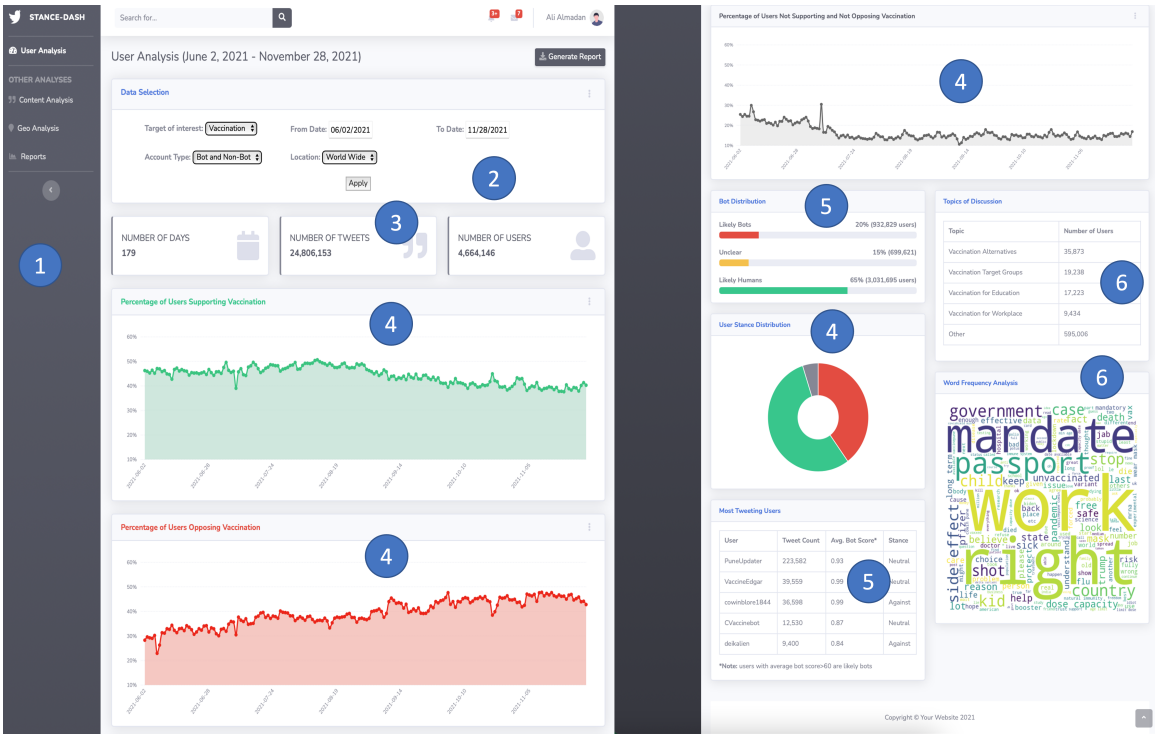

Figure 6.1: Screenshots of StanceDash and its components

6.4.2.1 Navigation Component

The navigation component (1) is presented as the navigation bar on the left. The navigation bar assists end users navigate through major analyses and reports. This includes content analysis, geo analysis, and reports. Although the navigation com-

[1]https://startbootstrap.com/theme/sb-admin-2

ponent has links to these other functions, the evaluation study in this research only covers the user analysis.

6.4.2.2 Data Selection Component

The data selection component (2) provides a utility for analyzing a partial view of the data. An end user can select the stance target, the date range, filter data by account type (bot and non-bot), and select a geographical location. To minimize and prevent errors, we used drop-down menus and date pickers. The advantage of using date pickers is that end users can select the date without the pardon of the date format. Therefore, it can be usable for users residing in different countries and not only in the United States.

6.4.2.3 Summative Data Component

The summative data component (3) provides simple, but important, numbers about the selected data. It shows the number of days, tweets, and users. In addition, most tweeting users along with their tweet count, bot score, and stance are presented in tabular data. This allows the end users to gain insights about the users involved in the discussion.

6.4.2.4 Stance Data Component

To achieve G1 and G2, we consider the stance data component (4) the most essential component in StanceDash. It provides end users with a tool to track public opinion on daily basis. The stance component has three charts: percentage of users supporting vaccination, opposing vaccination, and not supporting or opposing vaccination. Presenting public opinion in Twitter data using three charts allows end users to understand the correlation between the three aspects of public opinion. For example, an increase in the percentage of users opposing vaccination does not necessarily indicate an increase in the percentage of users supporting vaccination, but it could be attributed to the increase in the percentage of users on neither side (neutral stance).

To summarize public opinion for a specific period of time, we used a pie chart that shows the percentage of users in each stance category. End users can hover over the parts in the pie chart to explore public opinion for the specified date range. In all charts, we used red, green, and gray colors to represent support, opposition, and neutral stances, respectively.

6.4.2.5 Bot Data Component

The purpose of the bot data component (5) is to provide in-depth analysis of the Twitter users that contribute to the discussion related to the stance target. The data can be filtered from the data selection component to include/exclude bots accounts from the displayed data. In addition, the bot scores for most tweeting users along with their stances are shown in a tabular data presentation.

To recap from Chapter 5, we used Botometer V4 to assess whether an account exhibits automated behavior. Botometer assigns bot scores between 0 and 1, where a score of 0 indicates that the account is unlikely to be a bot account and a score of 1 indicates that the account is likely to be a bot account. Botometer has English and universal scores, where the former incorporates language-specific features to determine the likelihood for an account to be a bot (bot score). Since our data consist of English tweets only, we only showed the English scores in the dashboard. Additionally, we included the interpretation of the bot scores.

6.4.2.6 Qualitative Data Component

An essential part of understanding public opinion is understanding the language and topics of discussion and concerns. Therefore, we performed topic modeling and included the results in StanceDash as a qualitative data component (6) to allow end users understand the topics that are related to stance-taking. To determine the topics in the data, we used BERTopic, a neural state-of-the-art topic model [117]. BERTopic was validated on 16,309 news articles and 44,253 tweets. In their evaluation, the

authors report that BERTopic had high topic coherence and diversity when used on tweets, outperforming traditional topic models such as LDA. We followed the authors' example[2] for preprocessing tweets and applied additional steps:

- Removing URLs from the tweets
- Removing mentions (@) from the tweets
- Removing hashtags (#) from the tweets
- Removing stop words ((e.g. he, she, it, can, etc.))
- Converting all words in the tweets to lowercase
- Lemmatizing all words in the tweets

After preprocessing the tweets, we applied BERTopic on tweets within the date ranges and included the results in StanceDash. We show an example of topics associated with vaccination discussion between August 21, 2021 and August 28, 2021 in Figure 6.2. The intertopic distance map is shown in Figure 6.3.

For the results of topic modeling, we manually inspected the words with the highest scores in each topic from the output of BERTopic and assigned labels to the most frequent 4 topics. In this example, we can see that *Topic 0* is related to infection prevention as an alternative to vaccination, *topic 1* is related to vaccination target groups, *topic 2* is related to education, and *topic 3* is related to the workforce. For each topic, we examined random tweets to improve our topic labels. In our dashboard, we show the 4 most frequent topics, and we combine the rest of the topics into one topic and named it *Other*. We can see from Figure 6.3 that the topics are well separated. However, they included sub-topics. In this research, we only consider the main topics.

[2]https://github.com/MaartenGr/BERTopic

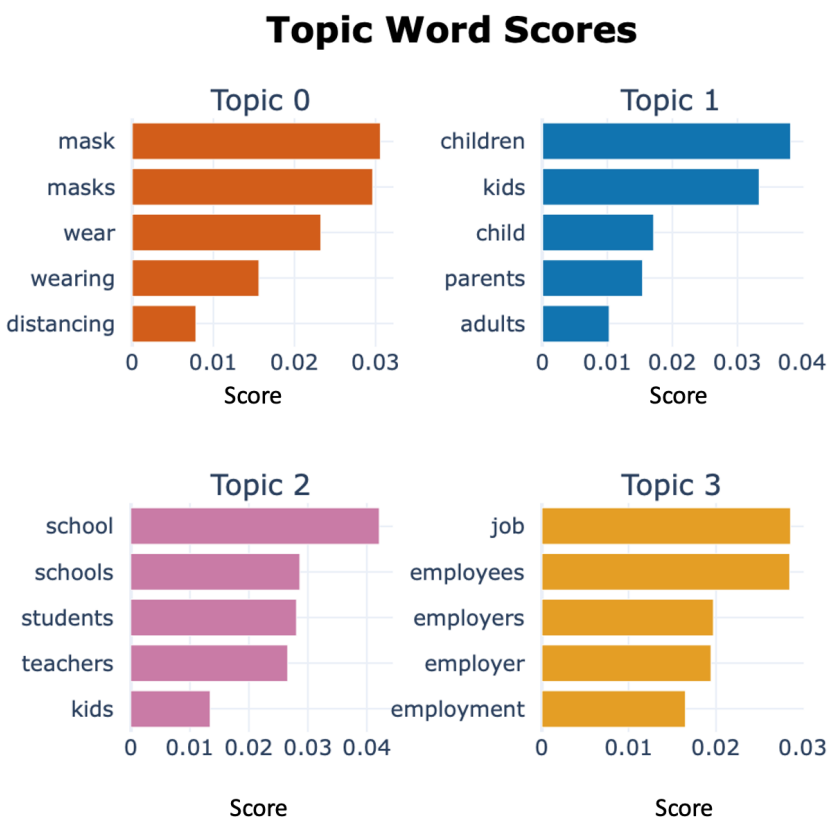

Figure 6.2: Topics of discussion between August 21, 2021 and August 28, 2021

The results from topic modeling provided us with the number of tweets for each topic. However, a user-based stance analysis requires understanding the number of users that discuss each topic. For this purpose, we inspected the tweets for each topic and maintained the number of users who contributed to each topic. Note that a single user could contribute to more than one topic. Therefore, the sum of users for all topics can exceed the number of unique users that contributed to the vaccination discussion during any period of time.

To complement topic modeling, we included a word cloud, a visual representation of the most frequent words in the data. Word clouds can serve as powerful summarizing tools. In learning settings, various studies suggest that exposure to word clouds can provoke critical thinking [118, 119]. The purpose of the word cloud in this research is to understand the underlying reasons for stance-taking. For example, the presence

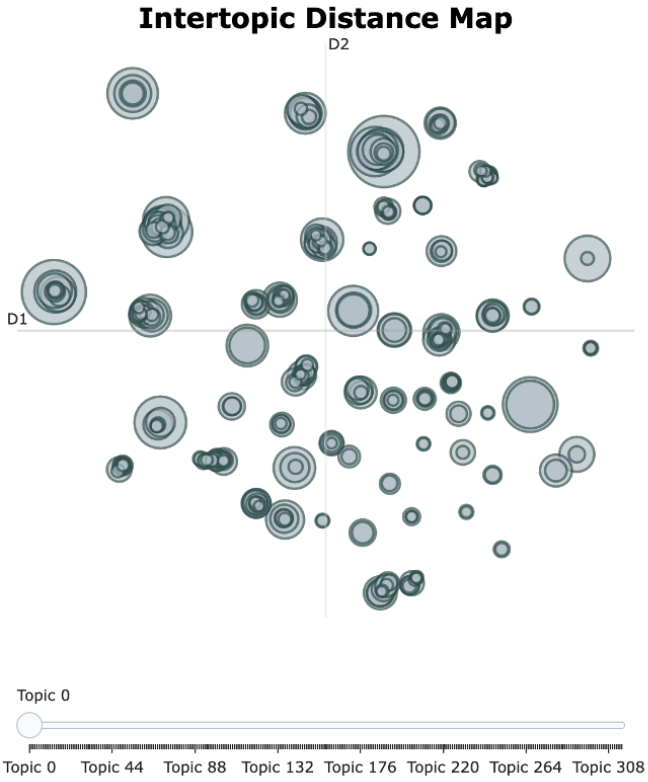

Figure 6.3: Inter-topic distance map for topics between August 21, 2021 and August 28, 2021

of the word *government* in the word can be indicative of a stance against vaccination due to conspiracy discussions. This can justify why Twitter users discuss vaccination alternatives, such as masks and social distancing, widely in the data.

6.5 Evaluation Study Design

This section explains the study design to evaluate StanceDash. The study had two specific goals:

1. Assess whether StanceDash has met its design goals (G1 and G2)
2. Identify any usability issues that could emerge from the interaction of end users with StanceDash

This study was approved by the Institutional Review Board (IRB) at the University

of North Carolina at Charlotte to protect human subjects. The associated record number is IRB-23-0008.

6.5.1 Recruitment and Participants

To recruit participants, we sent a recruitment email on October 4, 2022 to undergraduate and graduate students at the University of North Carolina at Charlotte through the university's listserv, an application that distributes messages to subscribers on an electronic mailing list. Our inclusion criteria specified that a potential participant must be 18 years or older and be able to participate in the study in person. The recruitment email included a Doodle signup link, where interested individuals could sign up for a time slot to participate in the study. Therefore, the recruitment of participants was conducted on a first-come-first-serve basis. Then we emailed the individuals who signed up for a time slot and shared the consent form with them on a Google form. Participants were asked to complete the consent form before coming to the study and provided contact information to answer any questions they had about the study and the consent form.

Although usability testing research suggests that 5 participants are sufficient to uncover most usability problems [120], we recruited 13 participants (8 male and 5 female participants) for this study to collect more comprehensive data to evaluate StanceDash. The average age of the participants was 27 years old, ranging from 19 to 35. The participants were fairly familiar with Twitter and experienced with interacting with data visualization tools. Table 6.1 shows the background and demographic information of the participants. Participants who completed the entire study received $10 Amazon Gift Cards. All participants completed the study in October 2022, but one participant (P11) was excluded from the analysis because he did complete the tasks according to the given instructions.

Table 6.1: Participants' background and demographic information

	Age	Gender	I am familiar with Twitter	I am experienced in interacting with data visualization tools
P1	32	Male	Strongly Agree	Agree
P2	28	Male	Agree	Strongly Agree
P3	35	Male	Agree	Neutral
P4	-	Male	Disagree	Agree
P5	25	Female	Strongly Agree	Agree
P6	31	Male	Agree	Strongly Agree
P7	19	Female	Strongly Agree	Agree
P8	26	Male	Agree	Agree
P9	19	Female	Neutral	Agree
P10	26	Female	Agree	Neutral
P11	34	Male	Strongly Agree	Neutral
P12	19	Male	Strongly Agree	Agree
P13	26	Female	Strongly Agree	Agree

6.5.2 Setup

The study was carried out in a controlled environment at a computer laboratory at the University of North Carolina at Charlotte. We used one computer device to present the dashboard to the participants and collect their responses. The 27-inch computer screen was split into two halves: The right half had StanceDash and the left half had tasks and questions on Qualtrics. We audio- and screen-recorded the interaction between the participants and StanceDash. We also video recorded the sessions to ensure that the participants were focused on the study and not distracted or interrupted.

6.5.3 Procedure

When a participant arrived, they were welcomed and seated in front of the computer screen with the dashboard and tasks on the screen. We then checked whether they had signed the consent form before starting the study. If they did not, they were asked to read and sign the form and given the opportunity to ask any questions. Before the start of the study, we gave the participants a 5-minute demonstration of

the dashboard's components and we answered any questions they had. We utilized the concurrent think-aloud protocol [121]. Therefore, we emphasized to the participants the importance of articulating their thoughts and feelings loudly while interacting with the dashboard and completing the tasks. Each evaluation study was dedicated 1 hour. The design and flow of the study are shown in Figure 6.4.

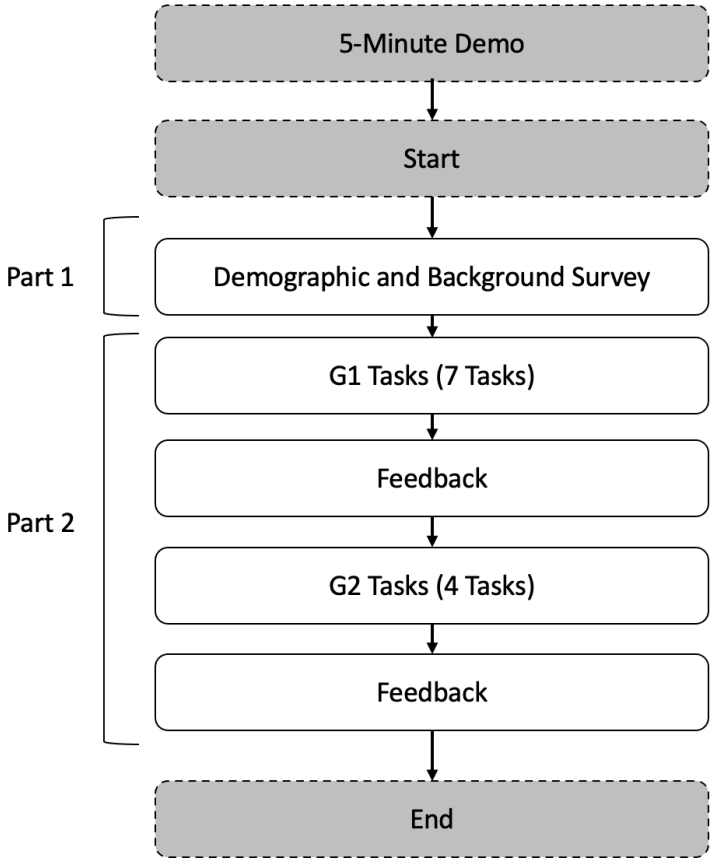

Figure 6.4: The design and flow of the evaluation study

We designed several tasks that required participants to interact with the dashboard. The tasks were carefully designed to align with the research question and the design goals (G1 and G2) of the dashboard. In the process of designing the tasks, we considered different levels of difficulty. We also designed tasks and questions to allow participants to interact with different menus and date ranges. We also designed questions to allow participants to provide their feedback. For the first design goal

(G1), we asked the participants to complete the following tasks:

- **T1:** Looking at the percentage of users that support and oppose vaccination, how would you describe the change of public opinion (stance) toward vaccination over time between June 2, 2021 and November 28, 2021 worldwide?
- **T2:** What was the percentage of all users (bots and non-bots) that were in support of vaccination on September 26, 2021 worldwide?
- **T3:** What was the stance of the user with the highest number of tweets worldwide between June 2, 2021 and August 1, 2021?
- **T4:** Which state (Texas vs. California) had a higher percentage of users (bot and non-bots) supporting vaccination on July 26, 2021?
- **T5:** How many bot users contributed to the vaccination discussion between September 1, 2021 and September 30, 2021 worldwide?
- **T6:** On which date did bots have the greatest support for vaccination between September 1, 2021 and September 30, 2021 worldwide?
- **T7:** What was the percentage of users that were 1) bots AND 2) opposing vaccination between September 1, 2021 and September 30, 2021 worldwide?

For the second design goal, we asked participants to answer the following questions:

- **T8:** What was the most discussed vaccination topic by all Twitter users (bots and non-bots) between June 2, 2021 and August 1, 2021 worldwide?
- **T9:** What were the vaccine-related topics that bots discussed between October 1, 2021 and November 28, 2021 worldwide?
- **T10:** What words did bots use the most worldwide between September 1, 2021 and September 30, 2021? Specify the top 3 words.

- **T11:** What topic did all Twitter users (bots and non-bots) discuss the most in the state of Texas between June 2, 2021 and October 1, 2021? How is the topic different from the most discussed topic in the state of California between June 2, 2021 and October 1, 2021?

After completing each of T7 and T11, participants were asked three questions to provide their comments and feedback on their interaction with StanceDash:

- **F1:** Completing this task, is there anything interesting that the dashboard helped you understand about measuring public opinion from Twitter data? If yes, please explain
- **F2:** In your opinion, were the information and visualizations provided on the dashboard helpful in achieving the task? Explain your answer
- **F3:** Do you have any suggestions to improve the dashboard and visualization to achieve the tasks? If yes, please explain

After the participants completed all tasks and answered all feedback questions, the study ended and we stopped recording. In the following section, we present our approach to evaluating StanceDash and the results.

6.6 StanceDash Evaluation

To evaluate the design goals and usability of StanceDash, we used mixed methods, an approach that uses collecting and analyzing quantitative and qualitative data within a single study [122]. The advantage of using mixed methods is that it allows for in-depth analysis and understanding of usability issues. In the first subsection, we explain our quantitative analysis and present our results. In the second subsection, we describe our qualitative analysis and our interpretation of the results.

6.6.1 Quantitative Analysis

First, we conducted a quantitative analysis of the data we collected to evaluate StanceDash. We mainly relied on the data we obtained from the screen and audio recordings of the interaction between the participants and the dashboard. Quantitative analysis is useful to assess usability. For this purpose, we used two quantitative usability metrics: completion time and success rate. The summary of the results for the two metrics is shown in Table 6.2.

The completion time is one of the widely used metrics to assess the usability of a system. In our evaluation study, the completion time CT for a task t is calculated by the average time the participants needed to complete the task, as shown in Equation 6.1:

$$CTt = \frac{Time_1 + Time_2 + ... + Time_n}{\text{The Total Number of Participants in Task t}} \tag{6.1}$$

We measured the completion time in seconds. Because there were tasks where participants did not need to interact with the data selection section, and because the typing speed of the participants varied, we measured the completion time from the moment the participant displayed the correct data to the time the participants completed the task verbally. The box plots for the completion time metric for all tasks are shown in Figure 6.5.

Box plots are advantageous for examining the spread of the data and detecting outliers. From the box plots in Figure 6.5, we found that there were outliers in the completion time metric for several tasks (T1, T3, T5, T6, T7 and T8). Therefore, we reviewed the videos and screen recordings to determine if there were external factors, such as technical issues, that affected the participants' ability to carry out the tasks. However, we could find any justifications that would warrant removing the outliers, and they were included in the analysis.

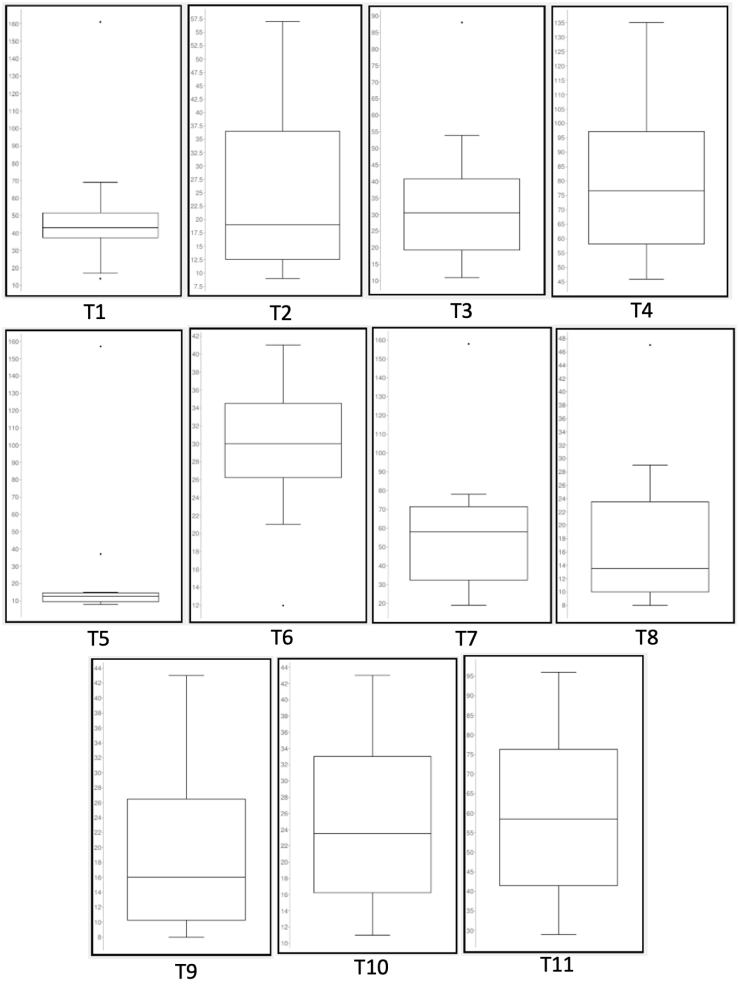

Figure 6.5: Box plots for the completion time metric for all tasks in the study

Analyzing the data, we observed that in tasks in which participants needed to compare public opinion in different locations (T4 and T11), participants took longer to complete these tasks compared to the other tasks. We anticipate that this long completion time is attributed to the fact that these two tasks required exploring and investigating two views in the data and required following extra steps. In T4, for example, the participants followed 4 steps to complete the task:

1. Displayed the data for one state.

2. Identified and memorized the percentage of users who supported vaccination on July 26, 2021 for the state they selected.

3. Displayed the data for the other state.

4. Identified the percentage of users who supported vaccination on 26 July 2021 and compared it with the other percentage from the other state.

Furthermore, some participants could not recall the percentage of users who supported vaccination in Texas when they explored the percentage in California. Therefore, they displayed the data for the state they first explored twice to ensure that they identified the correct percentages. We consider this to be a usability issue because participants needed to remember values to complete the task.

The second quantitative metric we used was the success rate. The success rate SR for a task t is calculated by the percentage of participants who were able to successfully complete task t with the expected outcome, as shown in Equation 6.2:

$$SRt = \frac{\text{The Number of Participants who successfully Completed Task t}}{\text{The Total Number of Participants in Task t}} \times 100 \quad (6.2)$$

In our study, we determined the successful completion by the participant's ability to complete the task correctly by providing the correct information in their responses. We found that the average success rate for the 11 tasks was approximately 88%, ranging from 75% to 100%.

Analyzing the results, we also observe that tasks T1 - T5 had a lower success rate, on average, than tasks T6 - T11. While the average success rate for T1 - T5 was 81%, the average success rate for T6 - T11 was 93%. An explanation of the improvement in the average success rate could be the presence of a learning effect. That is, the participants were not very familiar with the dashboard when they started the study. Overtime, they learned how to interact with it and where to find the required data and visualizations to complete the tasks.

Table 6.2: A summary of the two quantitative usability metrics we used to evaluate StanceDash

	Avg. Completion Time	Complete Success Rate
T1	50.8 seconds	83%
T2	26.9 seconds	83%
T3	33.8 seconds	83%
T4	**80.9 seconds**	**75%**
T5	25.9 seconds	83%
T6	29.1 seconds	100%
T7	58.5 seconds	92%
T8	17.6 seconds	100%
T9	19.7 seconds	**75%**
T10	24.6 seconds	100%
T11	**60.9 seconds**	92%

6.6.2 Qualitative Analysis

Next, we conducted a qualitative analysis. Our qualitative analysis was based mainly on the transcription of the audio recordings and the feedback questions that the participants responded to at the end of each of the two main tasks (F1, F2 and F3). Qualitative analysis is most useful to assess whether StanceDash achieved its design goals by allowing end users to measure, analyze, and monitor public opinion through a user-based stance analysis of Twitter data. Although audio recordings were valuable for evaluating moments of confusion and frustration, feedback questions allowed participants to provide comments on the dashboard visualization and strengths/weaknesses. We performed a thematic analysis [123] to identify the themes in the feedback from the participants and the think-aloud protocol. As a result, we identified 4 overarching themes that include 10 themes as shown in Figure 6.6. The data was coded by an expert coder.

6.6.2.1 Clarity and Usefulness

In general, participants positively commented on the clarity and usefulness of the dashboard. Most of the participants thought that the visualization of the data and

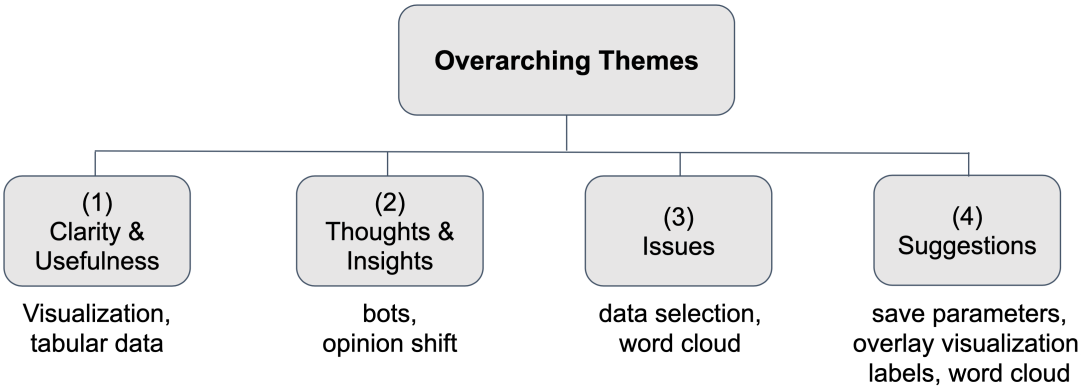

Figure 6.6: Qualitative analysis themes

the tabular data were easy to interpret and understand. They also indicated that they were helpful in achieving the tasks:

> *"They [visualizations] were helpful. The user distribution was indicative of the overall break down of the two opinions. It served as good summary tool"* [P4]

> *"The provided tables and graphics were helpful when answering questions about public opinion."* [P12]

However, we noticed from the screen recordings for the tasks that included identifying percentages of public opinion on a specific day that the participants were uncertain about which chart they should use to complete the tasks. In some instances, the participants completed the task by examining the percentages in the pie chart (user-stance distribution), while in fact this chart does not provide a percentage on a specific day, but rather for a whole period of time.

6.6.2.2 Thoughts and Insights

This overarching theme included comments from participants on interesting findings and thoughts they expressed while interacting with the dashboard and answering feedback questions. The participants indicated that the dashboard helped them

understand interesting things about measuring public opinion in Twitter data. Interestingly, most of the participants provided comments about the types of Twitter accounts. More specifically, they acknowledged the bots' role in shaping public opinion in Twitter data:

> *"There are a lot of bots on twitter. It probably makes it difficult to parse through information when they are so prolific. Every single most tweeting user I saw was likely to be a bot"* [P4]

> *"I would say that it is interesting to see how many bots are on the internet, and how much they can possibly sway on public opinion. I feel like this page should be on Twitters homepage so that people can be intelligent users of data"* [P8]

Other participants mentioned that the dashboard helped them understand the change of public opinion:

> *" the dashboard helped me to get meaningful insight about the change of public opinion about vaccination"* [P3]

> *" This dashboard is very helpful to gain insights of what people are thinking about vaccination. As after some time period people are opposing vaccination we can try to find out reason for that."* [P5]

It is evident from the participants' comments that the user-based stance analysis, which included the account type (bots and non-bots) provided them with insights about who is taking a stance in Twitter and the magnitude of bot involvement in the discussion. We believe that these thoughts and insights would not be observed if the dashboard provided a tweet-based stance analysis to measure public opinion from Twitter data.

6.6.2.3 Issues

In this overarching theme, the participants commented on some problems and difficulties while interacting with the dashboard and completing the tasks. We found that one usability issue in the StanceDash design is the data selection section. All 12 participants commented on the efficiency of the data selection section in the dashboard. The current design of StanceDash resets the menus and dates after the users click on the *Apply* button. Although the dashboard indicates in the top of the page the current data selection parameters, some participants indicated that they found it difficult to remember whether they used the correct parameters to filter and display the correct data view:

> *"As for some instance I was confused whether I applied that filter or not"* [P5]

> *"one thing that I mentioned this would immediately annoys me is that when I hit the filter, I kind of get paranoid I want to make sure that I got the right settings but it resets so then I'm not sure if I entered things incorrectly"* [P6]

Furthermore, P6 and P11 sighed while interacting with data selection as a way of conveying annoyance or dislike of the data selection section. Other participants stated that they could have performed the tasks more efficiently if the data selection parameters had been saved, especially for T4 and T11 where participants were asked to compare public opinion in two states:

> *"Now I have to put the dates again. If it wouldn't change it I wouldn't have to change the date again because it is the same time period"* [P10]

Other participants interacted with the data selection section, although the data they needed to interact with the data to achieve the task were already displayed. The

participants also commented on data visualization. Most of the participants indicated that the visualizations in StanceDash were helpful in completing the tasks.

Because participants were asked to identify the top 3 words the bots used in T10, all participants successfully completed the task. However, some participants indicated that they found it difficult to determine which word in the top 3 words was more frequent than the others.

> *"I was a little confused while deciding the most frequent words as two words looked the same size and I am not sure how to know which one is the most frequent one or if there is any axis related info that I missed."* [P10]

> *"The word frequency analysis isn't the most straight forward to understand. I think maybe a table listing the top words, similar to topics, would be more helpful."* [P12]

6.6.2.4 Suggestions

This overarching theme included suggestions from participants to improve the usability of the dashboard. Because most of the participants disliked the data selection section, they provided suggestions to improve its efficiency. They suggested that it would be helpful if the dashboard remembers the previous parameters to avoid entering the same information repeatedly.

> *"Save the inputed data for the data selection tool, so I can continue editing my parameters."* [P4]

To improve visualization, P1 and P7 suggested adding labels that show the percentage of users in each stance category in the pie chart without the need to hover over the section. In addition, P5 and P8 suggested adding the count of words to the frequency analysis (i.e., word cloud) to analyze the most frequent words that were used by Twitter users.

"Word frequency analysis can also display numbers of it." [P5]

"have a number of times associated with the word frequency." [P8]

6.6.3 Findings

Based on the quantitative and qualitative analysis in the evaluation study, we highlight our key findings. First, we found that StanceDash met its design goals to assist end users measure, analyze and monitor public opinion through a user-based stance analysis of Twitter data. Participants in the evaluation study were able to gain insight into public opinion from Twitter data for various locations and account types. For example, the participants were able to understand the magnitude of the bots' presence in the data. Additionally, they were able to compare the public opinion of Twitter users across different types and locations. Furthermore, some participants identified the role bots played in the discussion by not taking a stance toward vaccination. We argue that these observations would not have been possible without a user-based stance analysis. However, we found that there were usability issues in our design of the dashboard. For example, the participants found that the data selection section was not efficient. In some cases, participants were uncertain about which visualizations to interact with to complete some of the tasks. However, usability issues did not prevent participants from carrying out the tasks. Rather, they resulted in longer completion times for the tasks.

6.7 Summary

In this study, we designed and evaluated StanceDash, a web-based dashboard that assists end users to analyze and monitor public opinion with a user-based stance analysis in Twitter data. To achieve its design goals, StanceDash has six components: navigation, data selection, summative data, stance data, bot data, and qualitative data components. We evaluated the dashboard by recruiting 13 participants. Our findings showed that StanceDash was helpful and easy to use to gain insight into

public opinion in Twitter data. However, we found some usability issues that affected the performance of the participants in completing the tasks efficiently. One usability issue in the current design of StanceDash is the data selection section, where users select the date range and apply some filters to explore the data. Additionally, the participants highlighted some areas for potential improvement. The next version of StanceDash should take into consideration these discovered usability issues, as well as implement the suggestions that were received from the participants before the dashboard can be re-evaluated.

CHAPTER 7: FUTURE WORK AND CONCLUSION

In this chapter, we emphasize the contributions of this dissertation research by summarizing the findings, highlighting the limitations, and providing direction for future research and conclusions.

7.1 Contributions

This research makes significant contributions to the fields of computer and social sciences. First, it introduces user-based stance analysis in Twitter data and its relevance to the concept of public opinion in Chapter 4. The research in Chapter 4 was accepted for publication in the proceedings of the Future of Information and Communication Conference (FICC 2023). Second, it shows that a user-based stance analysis in Twitter data can mitigate the impact of social bots on the measured public opinion in Twitter data with stance detection in Chapter 5. The content of Chapter 5 was published in the proceedings of the International Conference on Social Informatics (SocInfo 2022) [124]. Third, it provides a design and evaluation of a web-based dashboard that assists end users measure, analyze, and monitor public opinion using a user-based stance analysis in Chapter 6.

7.2 Limitations

Although we showed that a user-based stance analysis can be effectively used to measure public opinion in Twitter data, we acknowledge that there are limitations to this dissertation research. The limitations are categorized into four categories: 1) limitations of using Twitter to measure public opinion, 2) limitations of considering original tweets only, 3) limitations of datasets and tools, and 4) limitations of the evaluation study. We explain each category in more detail in the following subsections.

7.2.1 Limitations of Using Twitter to Measure Public Opinion

In this dissertation research, we leveraged Twitter data to measure public opinion with stance detection. However, it is important to acknowledge that there are limitations of this approach. For example, one limitation of using Twitter data as a basis for gauging public opinion includes the characteristics of the sample of the population that uses Twitter. Therefore. Twitter users' population could be different from the population sample in other public opinion methods, such as polls and surveys. In a recent study in 2015 [125], researchers investigated the political representativeness of Twitter users in the 2012 presidential elections in the United States. The study findings showed that Twitter users were not representative of the age group of the voting population. For example, social media users were generally younger than the voting population [126]. Similarly, another study found that Twitter users in the UK have different age, gender, and education characteristics than the general population for political attitudes during the British election in 2015 [127]. Therefore, we emphasize the importance of understanding the representativeness of Twitter users for the stance target when measuring public opinion with Twitter data.

7.2.2 Limitations of Considering Original Text Tweets Only

In this dissertation research, we excluded retweets and quotes from our stance and bot analyses in Chapters 4 and 5. Although retweets might indicate that the user is supporting another opinion by taking the action of retweeting, our focus was only on original tweets and replies. We also excluded tweets with links to external websites, as coding the tweet requires manual inspection of the content of the external link. Tweets with images and videos were also excluded as they were not part of text classification and were outside the scope of our study. Our decision to exclude these tweets was based on the challenges associated with identifying the stance in these tweets.

7.2.3 Limitations of Datasets and Tools

In this dissertation research, we streamed tweets related to the vaccination discussion on Twitter for our stance and bot analyses in Chapters 4 and 5. Therefore, we only used one dataset with one stance target to detect the stance and obtain the results. The decision to use one dataset was due to the lack of the large datasets that could enable temporal analysis and restrictions of tweet dehydration. One limitation of using only one dataset with one target of interest is that the results may not be generalized to other data sets and other targets. However, the methods we used in these chapters generalize to other datasets and targets.

Furthermore, we use Botometer V4 in Chapter 5 to detect bots and their stance in our data. Botometer has limitations such as the tendency to assign high bot scores to accounts with low tweets counts. Therefore, we expect that Botometer misclassified some Twitter accounts. Moreover, to assign binary classes to accounts (bot and non-bot) and because Botometer provides a likelihood of an account to be a bot, we selected a threshold. In our effort to mitigate the impact of misclassification, we discarded accounts with bot scores in the middle between 0.4 and 0.6. This decision could have affected the results if a large number of accounts had bot scores within this specified date range.

7.2.4 Limitations of The Evaluation Study

To evaluate StanceDash, we utilized a concurrent think-aloud protocol in which we asked participants to express their thoughts and feelings loudly while interacting with the dashboard and completing the tasks. Therefore, it is possible that this approach had an impact on our quantitative analysis. Although we started measuring the completion time for each task from the moment the participants looked at the data to the time they verbally completed the task, some research showed that thinking aloud could have influenced the way participants approached tasks and tasks completion

times [121].

7.3 Future Work

In this section, we provide potential directions for future work. Our future work is motivated by the limitations of the current dissertation. In addition, we identify directions for future research based on the findings from this research.

We acknowledge that different datasets and stance targets can have distinct characteristics. For example, bots can play different roles in Twitter discussions. While in Chapter 5 we found that bots mostly had a neutral stance and were guiding users to locations where the vaccine was available, the role of bots could be different during presidential elections. In some cases, bots might not be involved in the discussion. In addition, our analysis was based on data collected using the streaming API. Additional research is needed to examine other data collected using the REST API and compare the results.

Next, we plan to investigate how a user-based stance analysis compares to public opinion as measured from traditional methods, such as surveys and polls, for a specific location (country or state). While we consider traditional surveys and social media data as two distinct measures of public opinion, a comparison between stance aggregation methods and surveys allows for the external validity of using stance analysis to measure public opinion and estimate the error.

Although we showed how a temporal analysis of the user-based stance analysis can show the aggregate public opinion change over time, one direction of future work is investigating the factors that cause an aggregate opinion change. We anticipate that one factor that causes aggregate public opinion change is that users enter and exit the discussion during a time period. That is, users who express a stance on a one day might not be present in the following days. Similarly, new users enter the discussion everyday to express their opinion. Another factor could be the fact that some users change their stance at some point in the discussion and cause a change in

the aggregate public opinion.

In this dissertation, we conducted a study to evaluate StanceDash and identify usability issues that the user might encounter. Our study revealed some usability issues, but showed that a user-based stance analysis in a dashboard can assist end users analyze and monitor public opinion in Twitter data. We plan to design a second version of the dashboard based on the results of the evaluation study in Chapter 7. Although for this initial version our participants were a sample of the general population, we plan to recruit participants from select groups such as journalists, researchers, and social scientists.

7.4 Conclusion

In this dissertation, we introduced user-based stance analysis to measure public opinion with stance detection in Twitter data. We addressed three research questions to better understand how a user-based stance analysis can be used effectively to measure public opinion in Twitter data. In the comparison between user-based and tweet-based stance analyses, we found that there is a statistical difference between the two measures. We acknowledged that bots impose a challenge to measure public opinion Twitter, but our analyses showed that, while a tweet-based stance analysis can impact the measure of public opinion, a user-based stance analysis is more robust to the presence of these accounts in the data.

We visualized the user-based stance analysis in StanceDash, and our study to evaluate the web-based dashboard that assists the end user monitor public opinion in Twitter data showed the ability of a user-based stance analysis to engage the participants in thinking about the different characteristics of users that express opinion in Twitter data, with minimal usability issues.

Although our analyses and findings were based on the stance of Twitter users toward vaccination in the era of COVID-19, additional research is required to assess the generalizability of the research findings to other datasets and stance targets. We

also anticipate that more functionality in StanceDash, such as content analysis, can provide additional effectiveness to enable end users to understand the different aspects of public opinion in Twitter data.

REFERENCES

[1] J. W. Du Bois, "The stance triangle," *Stancetaking in discourse: Subjectivity, evaluation, interaction*, vol. 164, no. 3, pp. 139–182, 2007.

[2] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proceedings of the international AAAI conference on web and social media*, vol. 5, pp. 450–453, 2011.

[3] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from twitter text," *Journal of Computational Science*, vol. 36, p. 101003, 2019.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[5] P.-F. Pai and C.-H. Liu, "Predicting vehicle sales by sentiment analysis of twitter data and stock market values," *IEEE Access*, vol. 6, pp. 57655–57662, 2018.

[6] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *International conference on social computing, behavioral-cultural modeling, and prediction*, pp. 231–238, Springer, 2012.

[7] P. Sobhani, S. Mohammad, and S. Kiritchenko, "Detecting stance in tweets and analyzing its interaction with sentiment," in *Proceedings of the fifth joint conference on lexical and computational semantics*, pp. 159–169, 2016.

[8] W. Magdy, K. Darwish, and I. Weber, "# failedrevolutions: Using twitter to study the antecedents of isis support," *arXiv preprint arXiv:1503.02401*, 2015.

[9] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "# isisisnotislam or# deportallmuslims? predicting unspoken views," in *Proceedings of the 8th ACM Conference on Web Science*, pp. 95–106, 2016.

[10] K. Darwish, W. Magdy, and T. Zanouda, "Trump vs. hillary: What went viral during the 2016 us presidential election," in *International conference on social informatics*, pp. 143–161, Springer, 2017.

[11] F. H. Allport, "Toward a science of public opinion," *Public opinion quarterly*, vol. 1, no. 1, pp. 7–23, 1937.

[12] A. Aldayel and W. Magdy, "Your stance is exposed! analysing possible factors for stance detection on social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–20, 2019.

[13] L. Borges, B. Martins, and P. Calado, "Combining similarity features and deep representation learning for stance detection in the context of checking fake news," *Journal of Data and Information Quality (JDIQ)*, vol. 11, no. 3, pp. 1–26, 2019.

[14] S. V. Vychegzhanin, E. V. Razova, and E. V. Kotelnikov, "What number of features is optimal: a new method based on approximation function for stance detection task," in *Proceedings of the 9th International Conference on Information Communication and Management*, pp. 43–47, 2019.

[15] B. W. Roper, "Public opinion surveys in legal proceedings," *American Bar Association Journal*, pp. 44–47, 1965.

[16] F. Wiseman, "Methodological bias in public opinion surveys," *The Public Opinion Quarterly*, vol. 36, no. 1, pp. 105–108, 1972.

[17] T. Lüke and M. Grosche, "What do i think about inclusive education? it depends on who is asking. experimental evidence for a social desirability bias in attitudes towards inclusion," *International Journal of Inclusive Education*, vol. 22, no. 1, pp. 38–53, 2018.

[18] T. J. DeMaio, "Social desirability and survey," *Surveying subjective phenomena*, vol. 2, p. 257, 1984.

[19] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: a literature review," *Quality & quantity*, vol. 47, no. 4, pp. 2025–2047, 2013.

[20] B. Liu *et al.*, "Sentiment analysis and subjectivity.," *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.

[21] J. M. Garcia-Garcia, V. M. Penichet, and M. D. Lozano, "Emotion detection: a technology review," in *Proceedings of the XVIII international conference on human computer interaction*, pp. 1–8, 2017.

[22] H. L. Childs, ""by public opinion i mean"," *The Public Opinion Quarterly*, vol. 3, no. 2, pp. 327–336, 1939.

[23] M. L. Atkinson, K. E. Coggins, J. A. Stimson, and F. R. Baumgartner, *The Dynamics of Public Opinion*. Elements in American Politics, Cambridge University Press, 2021.

[24] A. Campbell, P. E. Converse, W. E. Miller, and D. E. Stokes, *The American Voter*. New York: Wiley, 1960.

[25] B. I. Page and R. Y. Shapiro, *The rational public: fifty years of trends in Americans' policy preferences.* Chicago: University of Chicago Press, 1992.

[26] J. A. Stimson, *Public Opinion in America: Moods, Cycles, and Swings*. Boulder, CO: Westview Press, 2nd ed., 1999.

[27] J. A. Stimson, *Tides of Consent: How Public Opinion Shapes American Politics*. New York: Cambridge University Press, 2004.

[28] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, pp. 178–185, 2010.

[29] J. C. A. D. Lopez, S. Collignon-Delmar, K. Benoit, and A. Matsuo, "Predicting the brexit vote by tracking and classifying public opinion using twitter data," *Statistics, Politics and Policy*, vol. 8, no. 1, pp. 85–104, 2017.

[30] A. Bovet, F. Morone, and H. A. Makse, "Validation of twitter opinion trends with national polling aggregates: Hillary clinton vs donald trump," *Scientific reports*, vol. 8, no. 1, pp. 1–16, 2018.

[31] D. Biber and E. Finegan, "Adverbial stance types in english," *Discourse processes*, vol. 11, no. 1, pp. 1–34, 1988.

[32] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, 2021.

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

[34] T. Kiemel, A. J. Elahi, and J. J. Jeka, "Identification of the plant for upright stance in humans: multiple movement patterns from a single neural strategy," *Journal of neurophysiology*, vol. 100, no. 6, pp. 3394–3406, 2008.

[35] W. Ferreira and A. Vlachos, "Emergent: a novel data-set for stance classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, ACL, 2016.

[36] M. Qiu, Y. Sim, N. A. Smith, and J. Jiang, "Modeling user arguments, interactions, and attributes for stance prediction in online debate forums," in *Proceedings of the 2015 SIAM international conference on data mining*, pp. 855–863, SIAM, 2015.

[37] K. Darwish, W. Magdy, and T. Zanouda, "Improved stance prediction in a user similarity feature space," in *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pp. 145–148, 2017.

[38] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "Stance detection in web and social media: a comparative study," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 75–87, Springer, 2019.

[39] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (cnn-lstm)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020.

[40] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski, "Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 845–854, 2019.

[41] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 31–41, 2016.

[42] A. Addawood, J. Schneider, and M. Bashir, "Stance classification of twitter debates: The encryption debate as a use case," in *Proceedings of the 8th international conference on Social Media & Society*, pp. 1–10, 2017.

[43] A. Almadan, M. L. Maher, F. B. Pereira, and Y. Guo, "Will you be vaccinated? a methodology for annotating and analyzing twitter data to measure the stance towards covid-19 vaccination," in *Future of Information and Communication Conference*, pp. 311–329, Springer, 2022.

[44] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu, "Identifying breakpoints in public opinion," in *Proceedings of the first workshop on social media analytics*, pp. 62–66, 2010.

[45] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, p. e12189, 2020.

[46] C. D. Broad, "Emotion and sentiment," *The Journal of Aesthetics and Art Criticism*, vol. 13, no. 2, pp. 203–214, 1954.

[47] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[48] A. H. Nababan, R. Mahendra, and I. Budi, "Twitter stance detection towards job creation bill," *Procedia Computer Science*, vol. 197, pp. 76–81, 2022.

[49] C. P. S. Kaunang, F. Amastini, and R. Mahendra, "Analyzing stance and topic of e-cigarette conversations on twitter: Case study in indonesia," in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0304–0310, IEEE, 2021.

[50] Y. Samih and K. Darwish, "A few topical tweets are enough for effective user stance detection," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2637–2646, 2021.

[51] K. Joseph, S. Shugars, R. Gallagher, J. Green, A. Q. Mathé, Z. An, and D. Lazer, "(mis) alignment between stance expressed in social media data and public opinion surveys," *arXiv preprint arXiv:2109.01762*, 2021.

[52] M. Klenner, D. Tuggener, and S. Clematide, "Stance detection in facebook posts of a german right-wing party," in *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 31–40, 2017.

[53] K. Shalini, M. Anand Kumar, and K. Soman, "Deep-learning-based stance detection for indian social media text," in *Emerging Research in Electronics, Computer Science and Technology*, pp. 57–67, Springer, 2019.

[54] F. Xu and V. Keelj, "Collective sentiment mining of microblogs in 24-hour stock price movement prediction," in *2014 IEEE 16th conference on business informatics*, vol. 2, pp. 60–67, IEEE, 2014.

[55] M. Taulé, M. A. Martí, F. M. Rangel, P. Rosso, C. Bosco, V. Patti, *et al.*, "Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017," in *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*, vol. 1881, pp. 157–177, CEUR-WS, 2017.

[56] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, R. Paolo, *et al.*, "Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets," *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pp. 1–10, 2020.

[57] P. Sobhani, D. Inkpen, and X. Zhu, "A dataset for multi-target stance detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 551–557, 2017.

[58] A. Aker, A. Zubiaga, K. Bontcheva, A. Kolliakou, R. Procter, and M. Liakata, "Stance classification in out-of-domain rumours: A case study around mental health disorders," in *International Conference on Social Informatics*, pp. 53–64, Springer, 2017.

[59] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. Kralj Novak, "Stance and influence of twitter users regarding the brexit referendum," *Computational social networks*, vol. 4, no. 1, pp. 1–25, 2017.

[60] C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier, "Will-they-won't-they: A very large dataset for stance detection on twitter," *arXiv preprint arXiv:2005.00388*, 2020.

[61] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai, and F. Tajariol, "The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement," *Ieee Access*, vol. 9, pp. 33203–33223, 2021.

[62] D. Cucinotta and M. Vanelli, "Who declares covid-19 a pandemic," *Acta Bio Medica: Atenei Parmensis*, vol. 91, no. 1, p. 157, 2020.

[63] G. Ada *et al.*, "The importance of vaccination," *Front Biosci*, vol. 12, pp. 1278–90, 2007.

[64] A. A. Dror, N. Eisenbach, S. Taiber, N. G. Morozov, M. Mizrachi, A. Zigron, S. Srouji, and E. Sela, "Vaccine hesitancy: the next challenge in the fight against covid-19," *European journal of epidemiology*, vol. 35, no. 8, pp. 775–779, 2020.

[65] G. Troiano and A. Nardi, "Vaccine hesitancy in the era of covid-19," *Public Health*, 2021.

[66] C. S. Wiysonge, D. Ndwandwe, J. Ryan, A. Jaca, O. Batouré, B.-P. M. Anya, and S. Cooper, "Vaccine hesitancy in the era of covid-19: could lessons from the past help in divining the future?," *Human vaccines & immunotherapeutics*, pp. 1–3, 2021.

[67] M. Sallam, "Covid-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates," *Vaccines*, vol. 9, no. 2, p. 160, 2021.

[68] R. K. Gupta, A. Vishwanath, and Y. Yang, "Covid-19 twitter dataset with latent topics, sentiments and emotions attributes," *arXiv preprint arXiv:2007.06954*, 2020.

[69] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalina, and G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration," *arXiv preprint arXiv:2004.03688*, 2020.

[70] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis," *IEEE Transactions on Computational Social Systems*, 2021.

[71] K. Gunaratne, E. A. Coomes, and H. Haghbayan, "Temporal trends in anti-vaccine discourse on twitter," *Vaccine*, vol. 37, no. 35, pp. 4867–4871, 2019.

[72] G. Muric, Y. Wu, E. Ferrara, *et al.*, "Covid-19 vaccine hesitancy on social media: Building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies," *JMIR public health and surveillance*, vol. 7, no. 11, p. e30642, 2021.

[73] R. Baly, M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov, "Integrating stance detection and fact checking in a unified corpus," *arXiv preprint arXiv:1804.08012*, 2018.

[74] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the fake news challenge stance detection task," *arXiv preprint arXiv:1707.03264*, 2017.

[75] P. Bourgonje, J. M. Schneider, and G. Rehm, "From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles," in *Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism*, pp. 84–89, 2017.

[76] J. E. Chung and E. Mustafaraj, "Can collective sentiment expressed on twitter predict political elections?," in *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

[77] L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media," in *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, pp. 1–8, 2012.

[78] H.-H. M. Lee and W. van Dolen, "Creative participation: Collective sentiment in online co-creation communities," *Information & Management*, vol. 52, no. 8, pp. 951–964, 2015.

[79] U. Grandi, A. Loreggia, F. Rossi, and V. Saraswat, "A borda count for collective sentiment analysis," *Annals of Mathematics and Artificial Intelligence*, vol. 77, no. 3, pp. 281–302, 2016.

[80] L. Tavoschi, F. Quattrone, E. D'Andrea, P. Ducange, M. Vabanesi, F. Marcelloni, and P. L. Lopalco, "Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from september 2016 to august 2017 in italy," *Human vaccines & immunotherapeutics*, vol. 16, no. 5, pp. 1062–1069, 2020.

[81] E. M. Cody, A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Climate change sentiment on twitter: An unsolicited public opinion poll," *PloS one*, vol. 10, no. 8, p. e0136092, 2015.

[82] E. M. Cody, A. J. Reagan, P. S. Dodds, and C. M. Danforth, "Public opinion polling with twitter," *arXiv preprint arXiv:1608.02024*, 2016.

[83] A. Karami, L. S. Bennett, and X. He, "Mining public opinion about economic issues: Twitter and the us presidential election," *International Journal of Strategic Decision Sciences (IJSDS)*, vol. 9, no. 1, pp. 18–28, 2018.

[84] F. Gupta and S. Singal, "Sentiment analysis of the demonitization of economy 2016 india, regionwise," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence*, pp. 693–696, IEEE, 2017.

[85] L.-A. Cotfas, C. Delcea, R. Gherai, and I. Roxin, "Unmasking people's opinions behind mask-wearing during covid-19 pandemic-a twitter stance analysis," *Symmetry*, vol. 13, no. 11, p. 1995, 2021.

[86] A. Bechini, P. Ducange, F. Marcelloni, and A. Renda, "Stance analysis of twitter users: the case of the vaccination topic in italy," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 131–139, 2020.

[87] S. Poddar, M. Mondal, J. Misra, N. Ganguly, and S. Ghosh, "Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users," *arXiv preprint arXiv:2111.10667*, 2021.

[88] J. R. Zaller *et al.*, *The nature and origins of mass opinion*. Cambridge university press, 1992.

[89] Y. L. Hsieh, S. Rak, G. K. SteelFisher, and S. Bauhoff, "Effect of the suspension of the j&j covid-19 vaccine on vaccine hesitancy in the united states," *Vaccine*, vol. 40, no. 3, pp. 424–427, 2022.

[90] E. Mahase, "Covid-19: Us suspends johnson and johnson vaccine rollout over blood clots," 2021.

[91] 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.

[92] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

[93] G. M. Sullivan and R. Feinn, "Using effect size-or why the p value is not enough," *Journal of graduate medical education*, vol. 4, no. 3, pp. 279–282, 2012.

[94] A. Aldayel and W. Magdy, "Assessing sentiment of the expressed stance on social media," in *International Conference on Social Informatics*, pp. 277–286, Springer, 2019.

[95] J. Tachaiya, A. Irani, K. M. Esterling, and M. Faloutsos, "Sentistance: quantifying the intertwined changes of sentiment and stance in response to an event in online forums," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 361–368, 2021.

[96] S. E. Bestvater and B. L. Monroe, "Sentiment is not stance: Target-aware opinion classification for political text analysis," *Political Analysis*, pp. 1–22, 2022.

[97] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[98] D. Stukal, S. Sanovich, R. Bonneau, and J. A. Tucker, "Detecting bots on russian political twitter," *Big data*, vol. 5, no. 4, pp. 310–324, 2017.

[99] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, "Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate," *American journal of public health*, vol. 108, no. 10, pp. 1378–1384, 2018.

[100] A. Bessi and E. Ferrara, "Social bots distort the 2016 us presidential election online discussion," *First monday*, vol. 21, no. 11-7, 2016.

[101] O. Boichak, S. Jackson, J. Hemsley, and S. Tanupabrungsun, "Automated diffusion? bots and their influence during the 2016 us presidential election," in *International conference on information*, pp. 17–26, Springer, 2018.

[102] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?," in *Proceedings of the 26th annual computer security applications conference*, pp. 21–30, 2010.

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

[104] M. Zhang, X. Qi, Z. Chen, and J. Liu, "Social bots' involvement in the covid-19 vaccine discussions on twitter," *International Journal of Environmental Research and Public Health*, vol. 19, no. 3, p. 1651, 2022.

[105] K.-C. Yang, C. Torres-Lugo, and F. Menczer, "Prevalence of low-credibility information on twitter during the covid-19 outbreak," *arXiv preprint arXiv:2004.14484*, 2020.

[106] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance detection with bidirectional conditional encoding," *arXiv preprint arXiv:1606.05464*, 2016.

[107] M. Mohtarami, R. Baly, J. Glass, P. Nakov, L. Màrquez, and A. Moschitti, "Automatic stance detection using end-to-end memory networks," *arXiv preprint arXiv:1804.07581*, 2018.

[108] Q. Zhang, E. Yilmaz, and S. Liang, "Ranking-based method for news stance detection," in *Companion Proceedings of the The Web Conference 2018*, pp. 41–42, 2018.

[109] Q. Sun, Z. Wang, S. Li, Q. Zhu, and G. Zhou, "Stance detection via sentiment information and neural network model," *Frontiers of Computer Science*, vol. 13, no. 1, pp. 127–138, 2019.

[110] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th international conference companion on world wide web*, pp. 273–274, 2016.

[111] S. Poddar, M. Mondal, J. Misra, N. Ganguly, and S. Ghosh, "Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 782–793, 2022.

[112] R. M. Martins, V. Simaki, K. Kucher, C. Paradis, A. Kerren, *et al.*, "Stancexplore: Visualization for the interactive exploration of stance in social media," in *Proceedings of the 2nd Workshop on Visualization for the Digital Humanities*, 2017.

[113] K. Kucher, R. M. Martins, C. Paradis, and A. Kerren, "Stancevis prime: visual analysis of sentiment and stance in social media texts," *Journal of Visualization*, vol. 23, no. 6, pp. 1015–1034, 2020.

[114] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher, "What do we talk about when we talk about dashboards?," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 682–692, 2018.

[115] Z. Qu and J. Hullman, "Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 468–477, 2017.

[116] O. M. Yigitbasioglu and O. Velcu, "A review of dashboards in performance management: Implications for design and research," *International Journal of Accounting Information Systems*, vol. 13, no. 1, pp. 41–59, 2012.

[117] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[118] A. DeNoyelles and B. Reyes-Foster, "Using word clouds in online discussions to support critical thinking and engagement.," *Online Learning*, vol. 19, no. 4, p. n4, 2015.

[119] B. M. Reyes-Foster and A. DeNoyelles, "Influence of word clouds on critical thinking in online discussions: A content analysis," *Journal of Teaching and Learning with Technology*, vol. 5, no. 1, pp. 16–32, 2016.

[120] J. Nielsen, "Why you only need to test with 5 users," 2000.

[121] K. A. Ericsson and H. A. Simon, *Protocol analysis: Verbal reports as data.* the MIT Press, 1984.

[122] A. Tashakkori, C. Teddlie, and C. B. Teddlie, *Mixed methodology: Combining qualitative and quantitative approaches*, vol. 46. sage, 1998.

[123] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[124] A. Almadan and M. L. Maher, "User-based stance analysis for mitigating the impact of social bots on measuring public opinion with stance detection in twitter," in *International Conference on Social Informatics*, pp. 381–388, Springer, 2022.

[125] P. Barberá and G. Rivero, "Understanding the political representativeness of twitter users," *Social Science Computer Review*, vol. 33, no. 6, pp. 712–729, 2015.

[126] P. Barberá, "How social media reduces mass political polarization. evidence from germany, spain, and the us," *Job Market Paper, New York University*, vol. 46, pp. 1–46, 2014.

[127] J. Mellon and C. Prosser, "Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users," *Research & Politics*, vol. 4, no. 3, p. 2053168017720008, 2017.

APPENDIX A: IRB Approval

To:	Ali Almadan Political Science & Public Admin
From:	Office of Research Protections and Integrity
RE:	Notice of Exemption with Limited Review Approval
Approval Date:	26-Sep-2022
Exemption Category:	2
Study #:	IRB-23-0008
Study Title:	Evaluation of Stance-Dash: A Dashboard for Measuring and Analyzing Public Opinion with Stance Detection from Twitter Data

This submission has been reviewed by the Office of Research Protections and Integrity (ORPI) and was determined to meet the Exempt category cited above under 45 CFR 46.104(d). In addition, this Exemption has received Limited Review by the IRB under 45 CFR 46.111(a)(7). This determination has no expiration or end date and is not subject to an annual continuing review. However, you are required to obtain IRB approval for all changes to any aspect of this study before they can be implemented and to comply with the Investigator Responsibilities detailed below.

Important Information:

1. Face masks are optional on UNC Charlotte's campus. This includes classrooms and other academic spaces. Researchers conducting HSR activities in other locations must continue to adhere to local and state requirements in the setting where the research is conducted.
2. Face masks are still required in healthcare settings. Researchers conducting HSR activities in these settings must continue to adhere to face coving requirements.
3. Organizations, institutions, agencies, businesses, etc. may have further site-specific requirements such as continuing to have a mask requirement, limiting access, and/or physical distancing. Researchers must adhere to all requirements mandated by the study site.

Your approved study documents are available online at Submission Page.

Figure A.1: A copy of the IRB approval to evaluate StanceDash

APPENDIX B: Recruitment Email for StanceDash Evaluation

Seeking Participants to Evaluate a Web-Based Dashboard - Earn an Amazon Gift Card!

We kindly request your participation in a study to evaluate a dashboard that assists end- users understand and analyze public opinion from Twitter data. This study is part of research conducted at the University of North Carolina at Charlotte in the College of Computing and Informatics, Software and Information Systems Department. The IRB study number is IRB-23-0008.

In the study, you will be presented with the dashboard and given two tasks. For each task, you will be required to answer several questions by interacting with the dashboard while verbally describing the steps you follow to answer the questions. The study will be screen and video recorded. The estimated time of the study is 1 hour and no preparation is needed. Please note that in-person participation is required. After completing the entire study, you will receive a $10 Amazon Gift Card.

If you would like to participate, please go to the following link and sign up for one available time slot. If no time slot is available, this indicates that we have received the maximum number of participants for the study. However, more time slots might become available if any of the recruited participants withdraws from the study. Once you sign up for a time slot, you will receive an email that has a link to the consent form. Please note that you need to sign the consent form before you come to the study. You can end your participation at any time. If you have any questions, please contact Ali Almadan (aalmadan@uncc.edu)

Signup link: https://doodle.com/bp/alialmadan/stancedash

Thank you!

Researcher:

Ali Almadan, Ph.D. student

Email: aalmadan@uncc.edu

Faculty advisor:

Mary Lou Maher, Ph.D.

Professor and Director, Center for Education Innovation and Research

Email: mmaher9@uncc.edu

Software and Information Systems, UNC Charlotte