# University of Padova

Department of Mathematics "Tullio Levi-Civita"

*Master Thesis in Cybersecurity*

# Unmasking the Orchestra: Quantifying the Impact of Coordinated Accounts in Online Debates

*Supervisor*
Prof. Alessandro Galeazzi
University of Padova

*Co-supervisor*
Prof. Mauro Conti
University of Padova

*Master Candidate*
*Mahta Amraji*

*Student ID*
2041263

*Academic Year*
2023/2024

To the women of Iran and their ongoing fight for life and freedom.

# Abstract

As the threat from coordinated accounts that seek to manipulate online discussions continues to grow, the availability of publicly accessible social media data offers an invaluable resource for analyzing and understanding the impact of such harmful activities. This data provides a detailed view of how coordinated efforts influence public discourse, allowing us to uncover patterns, detect malicious behavior, and assess the broader implications of these tactics. In this thesis, we investigate the impact of coordinated accounts on online debates by analyzing Twitter discussions surrounding the 2015 COP21 and 2021 COP26 climate change conferences.

Current approaches for detecting coordinated accounts often lack fine-grained metrics to quantify their impact on online discourse over time. To address this, we propose an integrated approach that evaluates a range of measurements, including user position in information cascades, action delay, descendant counts, hashtag co-occurrences, and user interaction metrics. By combining these measurements, we develop a comprehensive set of methods to identify and assess the evolving impact of coordinated accounts in online debates.

This approach improves our understanding of how these accounts influence public debates and contribute to polarization, informing the design of systems, algorithms, and policies to mitigate these threats. Through a series of temporal analyses, we measure changes and trends in coordination rate, the relationship between coordination and polarization, and the evolution of debate diversity and toxicity. Additionally, we compare these dynamics between COP21 and COP26, and explore shifts in skepticism among the groups.

# Contents

# Listing of figures

# Listing of tables

# Listing of acronyms

**NLP** . . . . . Natural Language Processing

**BERT** . . . . . Bidirectional Encoder Representations from Transformers

**SMAs** . . . . . Social Media Activities

**PCA** . . . . . Principal Component Analysis

**LSTM** . . . . . Long Short-Term Memory

**LDA** . . . . . Latent Dirichlet Allocation

**SVD** . . . . . Singular Value Decomposition

# 1 Introduction

In the digital era, social media platforms are integral tools for communication, information sharing, and community building. Platforms such as Twitter, Facebook, and Instagram, which together account for billions of users across the world, serve as digital public spaces where people interact, share stories, and debate important issues. Social media has given us potential for communication and interaction that we did not previously possess, and we need to clearly establish what those potentials are[1]. Social media's ability to provide wider access to information has significantly altered how people consume news and information[2].

Nonetheless, the impact of coordinated accounts has grown to be a major concern in this vibrant digital ecosystem[2]. Pacheco et al.[3] discussed the effects of coordinated accounts on social media platforms, highlighting their potential to manipulate public opinion, amplify misinformation, and drive polarization. Coordinated campaigns, orchestrated through networks with competing interests, may seek to divide society by artificially elevating particular viewpoints and voices on social media platforms[4][5]. As a result, the reach and visibility of certain stories are deliberately inflated, possibly distorting conversations, reinforcing perceived truth in false narratives, and threatening democratic processes that rely on a well-informed public[5][6].

The growing prevalence of coordinated influence efforts on social media presents unique challenges for the digital age, building on past methods of public manipulation that were often channeled through mainstream media outlets and centralized power.

Novel computational methods, such as machine learning algorithms and network analysis techniques, open up opportunities for automatically identifying and tracking coordinated ac-

tivities across social media platforms[7]. By analyzing collective user behavior, communication patterns, or information diffusion, it is possible to develop new ways to understand these potentially manipulative forms and their impact on online discourse [8][9]. In this context, it is crucial to better understand the mechanisms behind these activities and their impact on shaping public opinion. Such understanding may help identify the factors that influence the resilience or susceptibility of the social media ecosystem—and, ultimately, our society—to information manipulation. Typically, the most vulnerable debates are those surrounding highly controversial topics, such as vaccinations, politics, or climate change, which are often targeted by coordinated disinformation campaigns.

In this thesis, we focus on studying the impact of coordinated accounts on social media discourse around and during one of the major global events, the climate conferences of COP21 and COP26[10].

The COP21 and COP26 climate change conferences are major milestones in the global effort to combat climate change, attracting significant attention from a diverse array of stakeholders, including governments, non-governmental organizations, activists, and the general public. These events are excellent case studies for investigating the influence of coordinated groups due to their high visibility and the contentious nature of climate change debates. Hristakieva et al.[5] emphasizes the importance of these events in understanding how climate-related discourse is manipulated and how it affects public perception and policy. Understanding the influence of coordinated accounts during major global events like COP21 and COP26 is crucial for several reasons. Firstly, these conferences attract heightened attention and engagement from individuals across the ideological spectrum, making them prime targets for coordinated operations. Secondly, the outcomes of these events have significant implications for global climate policy, which means that any distortion in public discourse can profoundly impact public perception and decision-making processes. Additionally, the recurring nature and clear structure of these events allow for a detailed analysis of how debates evolve and how coordinated behavior adapts over time. By analyzing activity patterns, post timing, and interaction networks, researchers can identify typical behaviors and develop strategies to detect and counteract such manipulative efforts. This understanding is vital for creating more resilient information ecosystems that can withstand and mitigate the impact of coordinated disinformation campaigns. Lastly, insights from studying the impact of coordinated accounts during COP21 and COP26 can guide the efforts to protect public discourse on social media platforms. As these platforms increasingly shape public opinion, maintaining the integrity and authenticity of online discussions is essential. By addressing the challenges posed by coordinated accounts,

researchers and policymakers can work towards creating a healthier and more informed digital public sphere[11].

Various methods are used to detect Coordinated Social Media Activities (CSAs), each with its own trade-offs and best practices. While conventional methods focused on the nature of the accounts as a way to distinguish between malicious and legitimate actions, more recent research moved the focus on the actions themselves and on the behaviors of the accounts, such as action delay, engagement, and using the same hashtags or keywords in a coordinated fashion[2]. However, while these methods are effective for detecting clear instances of coordination, they often lack the granularity needed to understand how coordinated accounts influence specific tweet characteristics and shape the broader discourse. To address this, our analysis not only identifies coordinated activities but also explores their impact on various aspects of social media posts, such as sentiment, toxicity, and engagement. Additionally, by comparing the dynamics between COP21 and COP26, we gain important insights into how these coordinated efforts evolve over time. To address these limitations, a more nuanced approach is required. This involves not only detecting coordinated actions but also analyzing their effects on public discourse. For instance, understanding how coordinated accounts influence polarization and toxicity within online discussions, as well as how these impacts vary across different topics and events, is crucial to developing effective strategies for mitigating the spread of misinformation and fostering healthier public discourse[12].

Furthermore, many current approaches focus primarily on the detection stage, without sufficiently addressing the subsequent analysis of the influence these accounts have on the evolution of the discourse. This gap highlights the need for more comprehensive approach to provide a thorough understanding of the impact of coordinated accounts[13]. Several studies have highlighted the importance of developing more sophisticated detection methods. Weber & Neumann[6] discuss how traditional detection methods fall short in capturing the full extent of coordination among malicious accounts. Hristakieva[5] further emphasizes the need for comprehensive approaches to dissect the complex interactions and strategies employed by these accounts.

This study addresses these limitations by proposing an integrated approach that benchmarks a variety of measurements. This approach combines several key metrics to provide a comprehensive set of methods for identifying and quantifying the influence of coordinated accounts. The metrics include: User Position in Information Cascades, Action Delay, Descendent Counts, Hashtag Co-occurrences, User Interactions.

To analyze the impact of coordinated accounts on polarization within online discourse, we

employed advanced sentiment analysis tools. Specifically, we used the syuzhet library in R and a pre-trained BERT model in Python, and compared the results of both methods.

For measuring toxicity and analyzing trends, we utilized the Perspective API in R alongside the pre-trained BERT model in Python, with results from both methods compared for validation.

Going beyond these metrics, we also measured skepticism and its trends across the two conferences, COP21 and COP26. Our analysis included examining correlations between coordination, polarization, and toxicity, as well as identifying patterns within these dynamics. This involves using sophisticated statistical methods and network analysis techniques to rigorously investigate how these elements interact and influence each other within the context of online discourse and how it changes over time.

By integrating these metrics, our approach not only detects coordinated accounts but also relates it with other debate characteristics such as toxicity and polarization. This multifaceted analysis provides a deeper understanding of how coordinated efforts influence public opinion and online discussion dynamics.

The rest of this thesis is organized as follows. Chapter 2 gives background that are necessary for a complete understanding. Chapter 3 defines the measures and methods considered and describes the steps. Chapter 4 provides the details of our research, and presents the corresponding results. Chapter 5 discusses more in detail the related works, and Chapter 6 concludes the thesis.

# 2

# Background

In this chapter, we provide the necessary background to understand the methodologies and approaches employed in this study. This chapter is structured to cover several key areas crucial for analyzing the interplay of coordinated accounts with different characteristics of social media debates. We begin with a general introduction to social media data, highlighting its relevance to significant global events such as COP21 and COP26. This section sets the stage for understanding the scope and importance of analyzing social media activities during these events. Next, we review existing methodologies for detecting coordination and discuss various analytical techniques employed in this field, including network analysis, behavioral analysis, and machine learning approaches. Following that, we give a background review of existing methods for analyzing polarization, followed by a discussion of various analytical techniques, including sentiment analysis, network analysis. We then explore the concept of toxicity in social media, emphasizing its significance and providing clear definitions. We also review existing methods for measuring toxicity, based on a thorough literature review. Finally, we define skepticism in social media, outlining its importance and providing definitions to frame our analysis. This section includes a literature review of existing methods for analyzing skepticism.

## 2.1 Social Media Analysis

### 2.1.1 Overview of Social Media Data

As of 2024, there are more than 5 billion active social media users around the globe and the average user spends 2 hours and 39 minutes on social media per day and these statistics are increasing every year[14][15]. Social media platforms have become vital channels for information dissemination, public engagement, and discourse. They generate vast amounts of data, encompassing a wide range of content, including text, images, videos, and user interactions. This data offers rich insights into public opinion, behavioral patterns, and the spread of information. In a more general manner, social media refers to Internet-based platforms for mass personal communication that facilitate interactions among users and derive their value primarily from user-generated content[16]. Also, social media data is characterized by its volume, velocity, and variety, making it a valuable resource for researchers. It allows for real-time analysis and the ability to track trends and changes over time. The complexity of this data requires advanced analytical methods to derive valuable insights[17], especially when analyzing aspects such as coordinated activities, toxicity, polarization, and skepticism.

### 2.1.2 Relevance to COP21 and COP26

Measures of public opinion and behavior are vitally important for shaping policy and improving our understanding of the social world. Traditionally, the most predictive and accurate method for social measurement has been sample surveys that ask carefully crafted questions to scientifically constructed samples of the population[18]. However, with the advent of social media, these traditional surveys are now complemented by the extensive and real-time data generated on various platforms. Social media offers a vast and diverse array of user-generated content, enabling researchers to capture a broader spectrum of public opinion and behavior on a much larger scale. This shift provides more immediate and nuanced insights into social dynamics, allowing for a deeper and more comprehensive understanding of public attitudes and trends[19].

This shift is particularly significant in the context of major global events like COP21 and COP26, where social media data provides an unparalleled opportunity to monitor and analyze public sentiment and discourse in real-time. Since social media has redefned the structure, dimensions, and complexity of the news[20][21]. By leveraging social media, researchers can gain deeper insights into how global climate conferences are perceived, the impact of coordinated

narratives, and the evolving dynamics of public opinion and engagement with climate change issues, all of which are crucial for shaping effective policies and understanding the broader social impact of these events[22]. Understanding the dynamics of social media discussions during COP21 and COP26 can shed light on broader patterns of information dissemination and public engagement. It can also reveal the tactics used by coordinated accounts to shape narratives, influence public opinion, and potentially sway the outcomes of such significant global discussions. This relevance underscores the importance of developing robust methodologies for analyzing social media data in the context of major international events.

## 2.2 Coordinated Accounts Identification

### 2.2.1 Definitions and Importance

Coordinated accounts on social media are accounts that act in a synchronized manner to manipulate discourse, spread misinformation, or amplify specific viewpoints. Coordination is defined as an unexpected, suspicious, or exceptional similarity among a set of accounts. It is often measured as the number of times two accounts behave similarly, such as when they repeatedly retweet the same posts[3][2][23].

Coordinated influence campaigns present a significant challenge to social media security. While automated bots can be identified through specific behavioral patterns, human-operated accounts engaged in coordinated activities can be more difficult to detect without comparative analysis[24]. Most existing methods for identifying and studying coordinated online behavior concentrate on isolated actions within specific social media platforms[25]. For instance, research has examined coordinated retweeting on X (formerly Twitter) by analyzing networks of accounts sharing identical content within short time-frames. Similar approaches have been applied to coordinated posting and link sharing on other platforms[26][27].

The study of coordinated accounts and their role in online information spreading can complement established studies on online information manipulation and inform the development of new and effective decision support systems against these manipulations.

Identifying coordinated accounts is crucial for several reasons. Firstly, it is essential for the preservation of information integrity. Coordinated accounts can spread misinformation and disinformation, which can mislead the public. By identifying and mitigating these accounts, the integrity of the information shared online is better preserved, ensuring that users receive accurate and reliable data[28].

Secondly, coordinated accounts can have a significant influence on public opinion. These accounts can amplify certain viewpoints, skewing public perception and influencing social and political outcomes. For instance, during elections, coordinated accounts can be used to promote specific narratives, impacting voter behavior and ultimately the election results[29].

Thirdly, identifying coordinated accounts is crucial for safeguarding democratic processes. These accounts can be used to manipulate voter behavior and influence election outcomes. Detecting and neutralizing such accounts is essential to protect the fairness and integrity of democratic elections and to prevent potential threats to the electoral system.[30].

Finally, recognizing coordinated accounts plays a crucial role in combating misinformation and fake news. By identifying these accounts, it is possible to reduce the spread of false information and ensure that public discourse is grounded in accurate facts[31]. This effort is crucial in maintaining a well-informed public and fostering a healthy, democratic society[31].

### 2.2.2 Existing Methods and Approaches

Traditional methods for detecting coordinated accounts often rely on identifying patterns such as high-frequency posting, the use of identical hashtags, or synchronized replies among multiple accounts. While these methods are straightforward and efficient in detecting obvious cases of coordination, they come with significant limitations. They tend to fail to uncover more sophisticated coordination strategies, as they focus primarily on detection without delving into the analysis of the accounts' influence on the broader discourse[32][33].

Traditional heuristic-based rules rely on predefined patterns to detect coordination. These methods are effective in identifying clear cases of coordinated behavior but fall short when it comes to subtle or evolving strategies. heuristic-based methods can be limited by their reliance on static rules, which may not adapt well to evolving strategies used by coordinated accounts[34]. This limitation highlights the need for more nuanced approaches like behavioral analysis, which focuses on the actions and interactions of users over time. Behavioral analysis can capture sophisticated and less obvious forms of coordination by examining posting frequency, timing, content similarity, and engagement patterns[35]. Combining heuristic-based rules with behavioral analysis can also enhance the detection of coordinated accounts. While heuristic-based rules can quickly identify clear cases of coordination, behavioral analysis can provide deeper insights into more sophisticated and evolving strategies6. This hybrid approach leverages the strengths of both methods, improving the overall effectiveness of detecting coordinated behavior[36].

One of the analytical techniques is network analysis; it examines the structure and dynamics of interactions among accounts to detect clusters of coordinated behavior. By mapping out the connections between users, we can identify unusual patterns that suggest coordination. Metrics such as community detection are often used to analyze these networks, providing insights into how coordinated accounts interact and influence the online space[36][35].

By integrating these advanced techniques, the identification and understanding of coordinated accounts can be significantly enhanced, providing deeper insights into their influence on public opinion, online discourse, and the broader landscape of misinformation and manipulation .

## 2.3 Polarization in Social Media

### 2.3.1 Definitions and Importance

Polarization in social media refers to the phenomenon where opinions, beliefs, or attitudes of individuals or groups become more extreme and divided over time[10]. This division often leads to the formation of distinct and opposing groups with little to no overlap in viewpoints, which can significantly impact public discourse and social harmony. In social media, polarization can be related to the phenomenon of echo chambers, where users are predominantly exposed to information that aligns with their existing beliefs. This reinforcement of views can lead to greater ideological divides. Whether echo chambers are a consequence of polarization or contribute to increasing the level of polarization is still a debated question[10][37][38][39]. However, research on polarization rates over time in social media reveals a complex and multifaceted relationship. Social media platforms, such as Twitter and Facebook, have been implicated in both increasing and decreasing political polarization[40]. Understanding and analyzing polarization is crucial for addressing its effects on society, particularly in the context of global interests such as climate change, vaccines, and politics.

Identifying and understanding polarization is essential because extremely polarized debate can have several negative effects on society. Firstly, polarization can hinder constructive dialogue and debate, leading to a fragmented and less informed public. When individuals or groups become deeply divided in their opinions, it becomes challenging to engage in meaningful conversations that bridge differing viewpoints, resulting in a public discourse that lacks depth and coherence[41][42].

Furthermore, polarized opinions can substantially influence social and political outcomes.

This influence extends to elections, policy-making, and social movements, where entrenched and opposing views can lead to radicalization or deadlocks[43][10].

Lastly, polarization can contribute to the spread of misinformation and fake news. Polarized groups are particularly susceptible to accepting information that aligns with their pre-existing beliefs, regardless of its accuracy[44]. This susceptibility exacerbates divisions and distorts public understanding of critical issues, further entrenching polarized viewpoints and making it difficult to achieve a well-informed, cohesive society.

### 2.3.2 Existing Methods and Approaches

Analyzing polarization in social media necessitates employing a range of methods and techniques aimed at measuring the extent of ideological division and uncovering the factors contributing to it [45]. Traditional and modern approaches are used to assess how user interactions and content contribute to polarized discourse. This section discusses key techniques including sentiment analysis and advanced network analysis methods that help in understanding polarization.

Sentiment analysis is a cornerstone technique for understanding the emotional tone of social media content [46]. By examining the sentiments expressed in posts, comments, and interactions, researchers can gauge the overall mood and attitudes of different user groups, offering valuable insights into how emotional tones contribute to polarization.

Early sentiment analysis tools, such as the syuzhet library in R, were designed to extract and quantify emotions from text. The syuzhet library employs various algorithms to analyze sentiment and map out the emotional arc of narratives[47]. This tool helps identify peaks and troughs in sentiment that may align with periods of polarized discussion.

In recent years, more advanced approaches such as BERT (Bidirectional Encoder Representations from Transformers) have revolutionized sentiment analysis. BERT, pre-trained on extensive text corpora and fine-tuned for specific tasks like sentiment classification, provides a nuanced understanding of context. Unlike traditional models, BERT considers both the left-to-right and right-to-left context, allowing for a deeper and more accurate sentiment classification into categories like positive, negative, and neutral [48]. This bidirectional context is particularly useful for interpreting complex social media content where understanding context is crucial for accurate sentiment analysis [49].

Beyond sentiment analysis, network analysis methods offer a valuable perspective on polarization. By constructing interaction networks from social media data, we can use node embed-

dings to derive ideological scores. For example, embedding methods like Node2Vec generate numerical representations of nodes (e.g., tweets or users) within a network, capturing their positional relationships and interactions[10]. A study measured ideology scores and using that the polarization by calculating ideology scores using latent ideology estimation. It involves creating a retweet matrix, filtering out users with limited retweet activity, and applying correspondence analysis with SVD; ideology scores are then derived from the SVD results and influencer scores are based on the median positions of users who retweeted them[10].

Combining sentiment analysis with ideology scoring provides a comprehensive view of polarization. By analyzing how sentiment correlates with ideological positions, we can further explore the interplay between emotional tones and ideological divides. This dual approach enhances our understanding of how different factors contribute to polarization in social media discussions.

Using these advanced tools and methodologies, we can gain a deeper insight into how emotional tones and ideological positions shape and reflect polarization in social media. This integrated approach helps in developing a more nuanced understanding of public discourse and its underlying dynamics.

## 2.4 Toxicity in Social Media

### 2.4.1 Definitions and Importance

Toxicity in social media refers to the presence of harmful, abusive, or offensive content that can degrade the quality of online discourse and negatively impact users' mental well-being[50]. This toxic behavior includes hate speech, harassment, threats, and other forms of harmful interactions that can lead to a hostile online environment[51]. Toxicity rates in social media are influenced by several factors, including the anonymity of users, the nature of online interactions, and the presence of inauthentic accounts. The lack of face-to-face cues and the anonymity provided by social media platforms can lead to more abrasive and confrontational interactions, increasing the salience of group memberships and generalizing negative experiences to entire outgroups[50]. Addressing toxicity is crucial for several reasons. Firstly, exposure to toxic content can lead to stress, anxiety, and other mental health issues for social media users. Creating a safer online environment is essential for promoting mental health and well-being. Secondly, toxic interactions can derail constructive conversations and lead to the spread of misinformation and polarization, thereby undermining the quality of public discourse. Lastly,

toxicity can exacerbate social divisions and contribute to real-world conflicts, making it essential to mitigate harmful behaviors online to ensure social and political stability[51][52]. Interestingly, toxic language does not necessarily discourage participation or escalate as discussions evolve[53]. However, exposure to toxic comments can increase the toxicity of subsequent comments[54].

### 2.4.2 Existing Methods and Approaches

Various methods have been developed to detect and manage toxicity on social media platforms, which can be broadly categorized into traditional approaches and advanced analytical techniques.

Traditional methods include keyword filtering and user reporting. Keyword filtering involves filtering content based on a predefined list of harmful keywords and phrases such as vulgar or offensive words or hate speech[50]. While easy to implement, this method can miss nuanced toxic behavior and generate false positives. User reporting and moderation rely on users to report toxic content, which is then reviewed by human moderators. This approach can be effective but is limited by scalability and potential biases in user reports[55][56].

A study by Avalle et al.[53] challenged the assumption that online platforms play a role in shaping toxicity by highlighting how human behavioral patterns remain consistent across platforms and eras. The researchers used Perspective API to analyze over 500 million comments. The research emphasizes the need for multifaceted approaches to moderation, as toxicity appears to be driven more by user sentiment and controversy.

NLP techniques analyze text to detect toxic content by examining syntactic and semantic patterns to understand the context and identify harmful language. Deep learning models, specifically LSTM networks combined with word embeddings like BERT, have shown promising results in toxicity classification, with accuracy reaching 94%[57]. Deep learning models like BERT, are trained on large datasets to recognize toxic behavior. These models can understand context and detect subtler forms of toxicity that keyword filtering might miss[7]. Pre-trained BERT models, fine-tuned for toxicity detection, offer state-of-the-art performance in identifying harmful content. These models consider the context and nuances of language, making them highly effective in recognizing complex forms of toxicity[58]. Effective toxicity detection often involves combining various methods. Furthermore, a widely used tool is the Perspective API developed by Jigsaw. This API uses machine learning models to assess the toxicity of text based on various attributes such as identity attack, insult, and profanity. It provides real-time

assessments that can be used to flag or filter harmful posts[59].

## 2.5 Skepticism in Social Media

### 2.5.1 Definitions and Importance

Skepticism in social media refers to the attitude of doubting or questioning the validity of information, events, or opinions shared on these platforms[60]. This skepticism can be directed towards various subjects, including news, scientific findings, political statements, and other public discourse[61]. Understanding and analyzing skepticism is essential for several reasons. First, it reflects the level of trust users have in the information they encounter online. High levels of skepticism can indicate a healthy critical engagement with content, but excessive skepticism can also lead to the dismissal of accurate information and the spread of misinformation[62]. Second, skepticism plays a crucial role in shaping public opinion and behavior, influencing how individuals perceive and respond to social, political, and scientific issues. Lastly, analyzing skepticism helps identify the factors that contribute to mistrust, which can inform strategies to improve information credibility and public trust in reliable sources[62][63]. These findings highlight the complex nature of online toxicity and its persistence across social media platforms.

### 2.5.2 Existing Methods and Approaches

Similar to toxicity, traditional methods include content analysis. Content analysis involves systematically examining the content of social media posts to identify expressions of doubt, questioning, or distrust. This method provides insights into the themes and subjects that elicit skeptical responses.

Advanced analytical techniques offer more sophisticated tools for analyzing skepticism. Machine learning approaches, particularly those involving NLP, are increasingly used to detect and analyze skepticism[60]. Sentiment analysis tools can be adapted to identify skeptical sentiment by training models to recognize linguistic cues associated with doubt and questioning. For instance, by analyzing the sentiment of posts, researchers can determine the overall mood and attitude of users towards specific topics, helping to identify skeptical viewpoints.

Topic modeling techniques, such as LDA, can uncover the main topics of skepticism by analyzing large datasets of social media posts[64][65]. These techniques help identify the subjects and themes that are most frequently questioned by users, providing a comprehensive under-

standing of the areas where skepticism is most prevalent.

In addition to these techniques, one of the most advanced approaches is the use of pre-trained BERT models. BERT has demonstrated high performance in understanding the context of words within a sentence, making it particularly effective for identifying nuanced skeptical language.

By fine-tuning a pre-trained BERT model on datasets containing examples of skeptical and non-skeptical content, the model can accurately classify new social media posts based on their skepticism levels. This approach improves the accuracy of skepticism detection by capturing the contextual subtleties of language. Additionally, BERT can be used to perform sentiment analysis, identifying emotions associated with skepticism. For example, skeptical posts often contain specific emotional tones, such as doubt, uncertainty, or distrust, which can be detected using BERT's advanced contextual understanding.

By employing deep learning methods and leveraging BERT for sentiment analysis and classification, researchers can gain a more accurate and nuanced understanding of skepticism on social media. These advanced techniques provide a powerful toolset for identifying skeptical sentiments and understanding the underlying emotions, themes, and topics that drive skepticism.

# 3
# Materials & Methods

In this chapter, we outline the data and methodologies used in this thesis. We begin by explaining the data collection process, which involves the aggregation of millions of tweets related to the COP21 and COP26 climate change conferences. Following this, we describe the data preprocessing steps undertaken to clean, transform, and filter the dataset, ensuring its relevance and quality for analysis. The chapter then continues with methods for the identification of coordinated accounts, a key aspect of our study, where we examine the propagation of tweets and the behavior of these accounts. We also present our approach to sentiment analysis, which uses advanced models to classify the emotional tone of tweets, and discuss our methods for analyzing toxicity, skepticism, and polarization within the online climate change debate. By exploring how these factors interact, we gain deeper insight into the dynamics of online discussions during these pivotal events. This chapter sets the foundation for the results and insights that follow in the subsequent sections of this dissertation.

## 3.1 Data

The data collection process for this thesis was focused on gathering a comprehensive dataset of Twitter posts related to the COP21 and COP26 climate change conferences. The objective was to create a robust and diverse dataset that would allow a thorough analysis of the discourse surrounding these two significant events and Twitter data is especially convenient for this purpose due to the rich structural data it captures[10].

In total, more than 5M tweets for COP21 and 10M tweets for COP26 between 2014 and 2021 were gathered as we can see the distribution in Figure?. The datasets includes various tweet attributes, such as tweet text, timestamp, user ID, retweet count, reply count, like count, quote count, and metadata such as hashtags, mentions, and URLs. This extensive datasets provides a rich source of information for analyzing trends, sentiments, and behaviors within the climate change debate in this thesis.

## 3.2 Identification of Coordinated Accounts

A critical aspect of our study is the identification of coordinated accounts within the Twitter dataset, particularly those involved in the diffusion of information related to the COP21 and COP26 climate change conferences. Coordinated accounts can potentially influence the spread of information, shaping public discourse through orchestrated actions such as retweets, likes, and replies. To identify these accounts, we used a multi-step process that involves analyzing the propagation patterns of tweets and the behavior of accounts within the dataset[2].

To enhance the robustness of our identification of coordinated accounts, we integrated three distinct measures: temporal action delay, network structure distance, and descendant count tracking[2]. By combining these measures, we created a comprehensive framework to identify patterns indicative of coordinated behavior. Temporal action delay allowed us to detect rapid retweeting activity that may signify organized efforts, while network structure distance provided insights into the propagation paths and connectivity among retweets[2]. Descendant count tracking helped us identify highly influential tweets and accounts that might play a central role in amplifying messages. To validate and refine our findings, we compared the accounts identified through these combined measures with those detected using the `CooRTweet` R package. The `CooRTweet` package, specifically designed to detect synchronized retweet patterns, provided an additional layer of analysis by highlighting clusters of coordinated activity based on retweet timing[66]. This comparative approach enabled us to cross-validate the trends in activities of coordinated accounts, detected using these measures, with our own. The integration of these methodologies and the subsequent comparison with `CooRTweet` results helped us to understand the role of coordinated accounts in shaping the discourse during the COP21 and COP26 conferences.

### 3.2.1 Root Account Identification

Identifying the root account is the first step in analyzing coordinated behavior on Twitter[2]. The root account is defined as the origin of a retweet cascade, the initial account that posted the original tweet that other accounts subsequently retweeted[2]. This process helps us trace the source of information spread and understand how it is propagated through the network.

#### Process for Identifying Root Accounts:

We employed text matching algorithms to find and group identical or near-identical tweets. By comparing tweet texts, we identified the original tweet that had been retweeted or quoted multiple times. Then we used the `created_at` timestamp to validate the chronological order of retweets and replies. The tweet with the earliest `created_at` timestamp in each cascade was considered the root tweet. This method ensures that the root tweet is accurately identified based on its original posting time. Finally, we constructed a network graph of retweets and replies to visually and analytically pinpoint the root tweet. By mapping out the interactions, we were able to trace the origin of each cascade more effectively.

### 3.2.2 Calculation of Distance from Root

After identifying the root tweets, the next step is to calculate the distance of each retweet from its original root tweet. This distance can be measured in terms of time and network structure and helps in understanding the dynamics of information propagation and potential coordination.

#### Temporal Distance

To analyze how quickly information spreads, we measured the temporal distance between the root tweet and each retweet. This process involves:

- **Sorting by Timestamp:** We used the `created_at` timestamp from the dataset to sort retweets chronologically. This allows us to determine the time intervals between the original tweet and each retweet.
- **Calculating Time Intervals:** We computed the time difference between the root tweet's `created_at` timestamp and the `created_at` timestamps of retweets. Short time intervals indicate faster propagation, which may suggest coordinated activity.

#### Network Structure Distance

In addition to temporal analysis, we examined the structure of retweet networks to understand how information spreads through the network. This involved:

- **Analyzing Retweet Paths:** We traced the propagation paths of retweets using various identifiers such as `retweeted_id`, `quoted_id`, and `in_reply_to_user_id`. This helped us map how retweets spread through the network.
- **Constructing Network Graphs:** We created network graphs to visualize the connections between root tweets and retweets. By analyzing these graphs, we identified clusters and patterns in the propagation of information.

By calculating the average distance of each retweet from its root tweet, we could identify propagation patterns that are indicative of coordinated behavior. On average, coordinated accounts are closer to the root of the information cascade[2].

### 3.2.3 Action Delay Measurement

To further understand the dynamics of coordinated behavior, we measured the time delay between the original root tweet and its subsequent retweets. The action delay is a critical metric, as coordinated accounts often exhibit rapid, synchronized retweeting behavior[2]. By analyzing these delays, we could identify clusters of accounts that retweeted within unusually short time-frames, suggesting coordination. This metric provided insight into the speed at which information was distributed by different accounts and helped to distinguish between normal and orchestrated retweeting patterns.

#### Process for Measuring Action Delay:

To measure action delay, we followed these steps:

- **Extract Timestamps:** For each retweet, we extracted the `created_at` timestamp from the dataset.
- **Calculate Delay Intervals:** We calculated the time delay for each retweet by subtracting the `created_at` timestamp of the root tweet from the `created_at` timestamp of the retweet. This results in the time interval between the root tweet and each retweet.
- **Analyze Distribution:** We analyzed the distribution of delay intervals to identify patterns. Short delays, particularly those within minutes of the root tweet, can indicate coordinated retweeting.

Through careful measurement and analysis of action delays, we gain a deeper understanding of the mechanisms behind information spread and potential coordination strategies during the COP21 and COP26 conferences.

### 3.2.4 Descendant Count Tracking

Another key aspect of our analysis was tracking the descendant count of each tweet[2], which includes metrics such as the number of retweets, replies, likes, and quotes. The descendant count provides a measure of the tweet's impact and reach within the network. By calculating the average descendant count for tweets from identified coordinated accounts, we could assess their influence on the overall topic. High descendant counts, particularly in coordinated networks, indicated that these accounts played a significant role in amplifying specific messages and narratives during the COP21 and COP26 events.

#### Process for Tracking Descendant Counts

To track descendant counts, we performed the following steps:

- **Data Extraction:** We extracted descendant count metrics from the dataset, including retweet counts, reply counts, like counts, and quote counts for each tweet. These metrics are provided in the dataset attributes.
- **Aggregation:** We aggregated these metrics to obtain a comprehensive view of each tweet's impact. For instance, the total descendant count for a tweet was calculated as the sum of retweets, replies, likes, and quotes.
- **Analysis:** We analyzed the aggregated descendant counts to identify tweets with high engagement. Tweets with the highest descendant counts were identified as highly influential.

### 3.2.5 Use of the `CooRTweet` Package

To go further with detection and analysis of coordinated behaviors, we employed the `CooRTweet` R package. This tool is designed specifically to identify and analyze synchronized retweet patterns on Twitter. The `CooRTweet` package offers several functionalities, including the detection of groups of accounts that retweet the same content within narrow time-frames, suggesting coordinated efforts[66]. It also provides metrics to quantify the level of coordination and visualization tools to map out networks of coordinated accounts.

Although CooRTweet relies solely on behavioral metrics, which differ from our approach, it was instrumental in identifying clusters of accounts that demonstrated coordinated retweeting behavior during COP21 and COP26. By comparing the results from CooRTweet with those obtained using our own method, we systematically assessed and verified the trends in coordinated activities over time.

Functionality of the CooRTweet Package:

The CooRTweet package is equipped with powerful tools for detecting synchronized retweeting behavior. It identifies groups of accounts that retweet the same content within narrow timeframes, which is a strong indicator of coordination. The package offers several key features:

- **Detection of Synchronized Retweets:** CooRTweet detects clusters of accounts that retweet the same tweet within a pre-defined short time interval. This is crucial for identifying potential coordinated campaigns.
- **Coordination Metrics:** The package provides various metrics to quantify the level of coordination, such as the number of retweets within specific time windows, and the number of unique accounts involved in these synchronized activities.
- **Visualization Tools:** CooRTweet includes visualization capabilities that allow us to map the network of coordinated accounts, illustrating how these accounts interact and amplify specific messages within the dataset.

## 3.3 Sentiment Analysis

The next step in this thesis was to perform sentiment analysis to further investigate the effect of coordination.

### 3.3.1 Syuzhet Package for Lexicon-Based Sentiment Analysis

We employed the Syuzhet[67] package, an R-based tool that uses lexicon-based approaches to extract sentiment from text. The Syuzhet package applies different sentiment lexicons, such as NRC, Bing, and AFINN, to provide a more granular understanding of the emotional undertones in tweets.

The lexicon-based sentiment analysis with the Syuzhet package involved the following steps:

- **Application of Sentiment Lexicons:** We applied multiple lexicons from the `Syuzhet` package to the tweet dataset. Each lexicon provided a different perspective on sentiment, ranging from basic positive/negative classifications to more nuanced emotional categories such as anger, joy, fear, and trust.
- **Comparative Sentiment Analysis:** By comparing the results from different lexicons, we could cross-validate the sentiment classification and identify any discrepancies. This comparative analysis added an additional layer of robustness to our sentiment analysis.
- **Temporal Sentiment Trends:** Using the timestamp data, we also analyzed how sentiment fluctuated over time during the COP21 and COP26 events. This temporal analysis helped us identify key moments when the sentiment shifted significantly, likely in response to specific events or announcements during the conferences.

### 3.3.2 BERT Model

We incorporated a BERT-based model to leverage the advantages of state-of-the-art machine learning techniques. This model was selected to enhance the accuracy and depth of our sentiment analysis, particularly given the multilingual and nuanced nature of the Twitter dataset.

#### Sentiment Scoring and Categorization:

The `nlptown/bert-base-multilingual-uncased-sentiment`[68] model, fine-tuned for sentiment analysis in multiple languages and case-insensitive. We assigned sentiment scores to the tweets on a scale from 1 to 5, where:

- **1 - Very Negative**
- **2 - Negative**
- **3 - Neutral**
- **4 - Positive**
- **5 - Very Positive**

To further streamline our analysis, we categorized the sentiment scores into three broader classes:

- **Positive:** Scores of 4 or 5.

- **Neutral:** Score of 3.
- **Negative:** Scores of 1 or 2.

This categorization allowed us to effectively distinguish between positive, neutral, and negative sentiments, simplifying the analysis and comparison across the dataset.

Emotion Classification:

In addition to sentiment scoring, we employed another BERT model specifically fine-tuned for emotion classification. This model was used to identify the predominant emotion in each tweet, providing deeper insight into the emotional tone of the discussions. The model was fine-tuned using the "bhadresh-savani/distilbert-base-uncased-emotion"[69] pre-trained model, which categorizes text into seven emotion labels: Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise.

This additional layer of analysis allowed us to go beyond simple sentiment analysis, offering a more detailed view of the emotional landscape during the COP21 and COP26 discussions. By combining sentiment scoring with emotion classification, we were able to capture a richer, more comprehensive picture of how these climate change events were perceived and discussed online.

## 3.4 Polarization Analysis

Following the sentiment analysis phase, we extended our investigation to explore the extent of polarization in the online discussions surrounding the COP21 and COP26 conferences. Polarization in this context is defined as the divergence of opinions towards extreme viewpoints[10][70], leading to a fragmented discourse where opposing perspectives have minimal common ground.

To understand the nature of polarization, we first examined the distribution of sentiment scores within our datasets. Sentiment analysis classified tweets into different emotional categories, ranging from very negative to very positive, with a neutral category included for balance. By analyzing the distribution of these sentiments, we aimed to identify the prevalence of extreme sentiments and the proportion of neutral ones and the presence of extreme sentiments were a sign of polarization in our case.

We employed network analysis techniques to derive ideological scores from social media interactions. Specifically, we constructed interaction networks for each conference using data on retweets, mentions, and replies.

For COP21, we built two types of networks: one based on retweets and another on mentions, and combined them to form a comprehensive interaction graph. Similarly, for COP26, we created retweet and mention networks and combined them. These networks were then analyzed using Node2Vec, a method that generates numerical embeddings for nodes (e.g., users or tweets) based on their network relationships. These embeddings were subjected to PCA to compute ideology scores.

The ideology scores derived from the embeddings were then analyzed to assess the distribution of ideological positions. We applied Hartigan's Dip Test to determine the modality of the ideology score distributions. This statistical test helps identify whether the distribution of scores is multimodal, which would indicate the presence of distinct ideological groups, which is a common sign of polarization.

We conducted a rolling correlation analysis at the end between the sentiment scores and ideology scores to further investigate the relationship between them in the discourse around COP21 and COP26. The rolling correlation plot displays how the relationship between ideology and sentiment fluctuates throughout the dataset. Peaks in the correlation may indicate times when ideological extremes are strongly associated with specific sentiment types, while troughs may suggest periods of less pronounced relationship or more neutral sentiment.

## 3.5 Toxicity and Skepticism Analysis

In this part of the thesis, we explored the analysis of toxicity and skepticism within the COP21 and COP26 Twitter datasets. Understanding the prevalence and distribution of toxic content and skepticism is crucial to evaluating the quality of discourse and the potential impact of harmful narratives on public opinion.

### 3.5.1 Toxicity Detection

To measure the toxicity of tweets, we employed two complementary approaches in R and Python:

#### BERT-Based Toxicity Model:

We used a BERT model specifically fine-tuned for detecting toxicity in text, known as `Unitary/toxic-bert`[71]. This model is adept at identifying and classifying toxic language, including insults, hate speech, and other forms of harmful content, in social media posts. The

BERT model's deep learning architecture allows it to understand the context and nuance of language, making it highly effective at distinguishing between benign and toxic expressions.

This model assigns a toxicity score to each tweet, where a higher score indicates a greater level of toxicity. The BERT model's ability to handle multiple languages and its deep learning architecture allow it to effectively capture the nuances of toxic language across different linguistic contexts.

Perspective API:

o further enhance our analysis, we leveraged the Perspective API, an R-based tool developed by Jigsaw and Google, which offers a comprehensive suite of toxicity-related attributes. The Perspective API provides scores for various dimensions of toxicity, including:

- `TOXICITY`: General likelihood that a comment is perceived as rude, disrespectful, or unreasonable, leading to a less civil discussion.
- `SEVERE_TOXICITY`: Measures the likelihood of extremely toxic comments that are likely to make someone leave a conversation.

### 3.5.2 Skepticism Detection

While the primary focus of this section is on toxicity, skepticism regarding climate change was also a critical factor in our analysis. Identifying skepticism involved detecting tweets that expressed doubt or disbelief in climate science or the need for climate action. Once again we employed a dual approach to identify skeptical tweets within the dataset:

Keyword-Based Identification:

The first method involved using a set of predefined keywords commonly associated with climate skepticism. These keywords included terms such as "hoax", "scam", "climate alarmism", "fake", "fraud", "lies", "myth", "swindle", "nonsense", and "conspiracy". Tweets containing any of these keywords were flagged as skeptical. This approach allowed us to quickly identify tweets that explicitly expressed doubt or disbelief in climate science or the urgency of climate action.

Sentiment Classification with BERT:

In addition to keyword detection, we tried to find a more accurate approach and applied sentiment classification using the BERT model monologg/bert-base-cased-goemotions-original[72] based on GoEmotions dataset[73]. This model was used to classify the emotional tone of tweets into 27 different categories such as ”anger”, ”fear”, and ”surprise”, etc. For skepticism detection, we focused on categories like ”confusion”, ”nervousness”, and ”fear” which are often associated with skeptical or oppositional sentiments toward climate change.

# 4

# Results

This chapter presents the results of the work done in this thesis. The analysis integrates multiple approaches to detect coordinated behavior, assess levels of polarization, toxicity, and climate change skepticism, and also tracks changes in trends across the two critical global events COP21 and COP26.

We begin by representing the results of our methodology that combines metrics such as the position in retweet cascades, action delay, and descendant counts to identify coordinated accounts. The detection algorithms were implemented using R and also using the CooRtweet package to validate our findings and enhance the reliability of our coordination detection even though CooRtweet uses different measures and it only considers behavioral techniques such as coordinated link or image sharing.

Additionally, for sentiment and toxicity analysis, we utilized advanced natural language processing techniques, implementing the BERT model in Python for ease of use and flexibility. This approach allowed for a robust analysis of how coordinated accounts influence the tone and content of the discussions.

Following the identification of coordinated accounts, the chapter explores in depth the effects these accounts have on the overall discourse. Specifically, we analyze how coordinated behavior correlates with increased polarization and toxicity in debates, and examine shifts in sentiment and the prevalence of skepticism.

A comparative analysis between COP21 and COP26 is also carried out to explore the evolution of these dynamics over time. This comparison highlights significant changes in the nature

and influence of coordinated accounts, as well as broader trends in public discourse on climate change.

So, the chapter is structured as follows: identification of coordinated accounts, their impact on discourse, detailed case studies, and a comprehensive analysis of sentiment, polarization, and toxicity.

## 4.1 Analysis of Coordination

We used combined metrics to detect coordinated accounts in Twitter debates around COP21 and COP26 by using several key metrics that reflect user behavior and network dynamics. The detection methodology integrates three primary metrics: position in retweet cascades, action delay, and descendant counts. Additionally, the results from this combined approach were cross-referenced with the outputs from the CooRtweet R package to validate the accuracy of the detected coordinated accounts.

### 4.1.1 Identification of Coordinated Accounts

The Results of the analysis is shown in the table 4.1 for COP21 and COP26. These accounts were flagged based on their consistent patterns across all three metrics. On average, coordinated accounts are closer to the root in retweet cascades, are faster in retweeting messages, and involve a higher number of downstream users (those who retweet or engage with the content further) compared to non-coordinated ones. These patterns suggest a coordinated effort to amplify content more effectively than non-coordinated accounts. we can see a slight increase in the coordination rate in COP26 event compared to COP21. This increase suggests a heightened effort by certain groups to manipulate public opinion and, therefore, potential evolution in the strategies employed by coordinated accounts over time, reflecting a more sophisticated approach to manipulating discourse.

| Event | Number of Tweets | Number of Accounts | Coordination Rate |
|---|---|---|---|
| COP21 | 5,712,050 | 935,573 | 11.5% |
| COP26 | 10,240,966 | 2,080,280 | 13.3% |

**Table 4.1:** Comparison of Coordination Rates

Table4.2 presents a comparative overview of the most frequently used hashtags and those specifically employed by coordinated accounts for the events, COP21 and COP26. The table

highlights the common hashtags observed in general discourse alongside the hashtags identified as being used by coordinated efforts to influence online conversations. By comparing these hashtags, the table provides insights into how hashtag usage evolved between the two events and underscores the role of coordinated accounts in amplifying certain topics or narratives within the online space. This comparison helps to understand the dynamics of hashtag influence and the strategic use of social media for public discourse manipulation.

| Event | Common Hashtags | Common Hashtags of Coordinated Accounts |
|---|---|---|
| COP21 | #COP21 | #COP21 |
| | #climatechange | #climatechange |
| | #climate | #climate |
| | #ParisAgreement | #Paris2015 |
| | #ClimateChange | #ClimateAction |
| COP26 | #COP26 | #COP26 |
| | #COP26Glasgow | #ClimateAction |
| | #ClimateAction | #ClimateCrisis |
| | #Glasgow | #ClimateEmergency |
| | #ClimateCrisis | #ecosocialism |

**Table 4.2:** Common Hashtags and Hashtags of Coordinated Accounts for COP21 and COP26

## 4.2 Sentiment & Polarization Analysis

### 4.2.1 Sentiment Analysis:

Sentiment analysis plays a significant role in understanding the emotional tone and public mood surrounding climate change discussions on social media. To perform this analysis, we first used the Syuzhet R library to classify public emotions and gain a better understanding of the general opinion about COP events. Then we utilized more advanced techniques, specifically leveraged the BERT model to deeply analyze and understand different aspects of the public discourse around the topic.

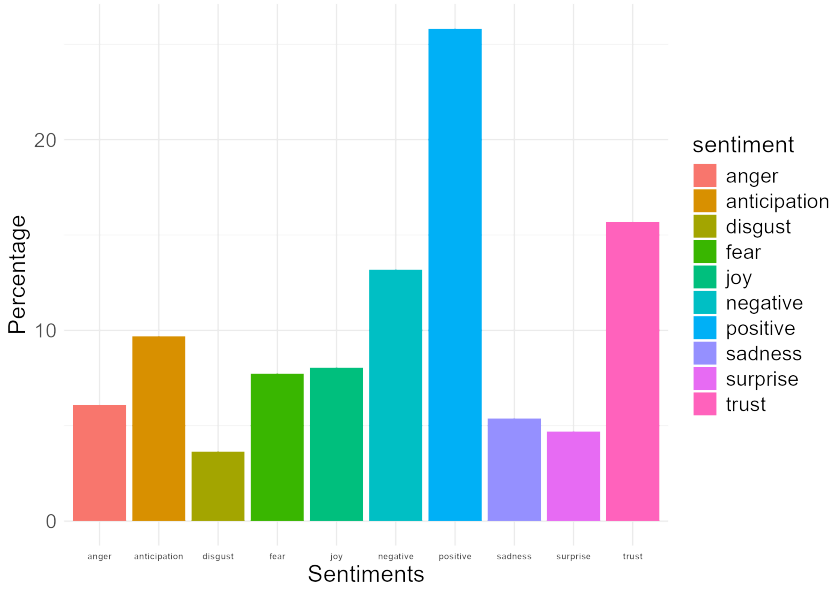


**Figure 4.1:** Result of Sentiment Analysis using Syuzhet for COP21

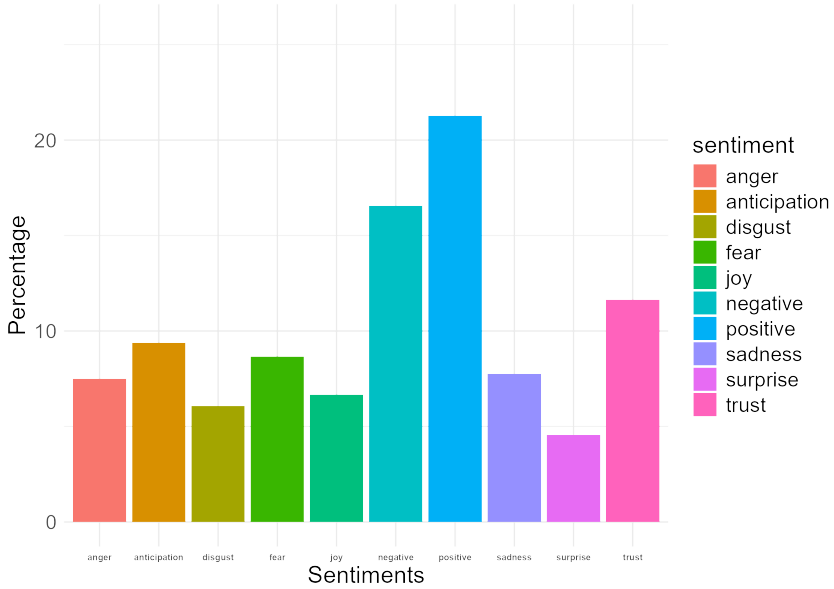


**Figure 4.2:** Result of Sentiment Analysis using Syuzhet for COP26

The graphs in Figure 4.1 and Figure 4.2 indicate the sentiment comparison between COP21 and COP26. It illustrates the frequency of various sentiments expressed, resulted from Syuzhet. The sentiments are categorized as Trust, Fear, Surprise, Sadness, Anticipation, Anger, Disgust, and Joy. Positive sentiments such as Trust and Anticipation were among the most frequent,

reflecting a hopeful and optimistic public mood. However, negative sentiments such as Fear and Anger were also more pronounced during COP26, indicating some public concerns and negative reactions. These findings underscore the increased emotional engagement and diverse public opinions surrounding COP26, offering valuable insights into the public discourse on climate change.

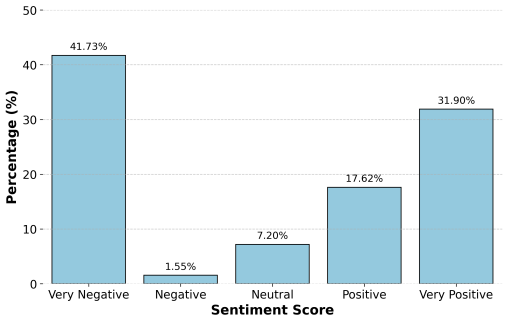


**Figure 4.3:** Result of Sentiment Analysis using BERT for COP21

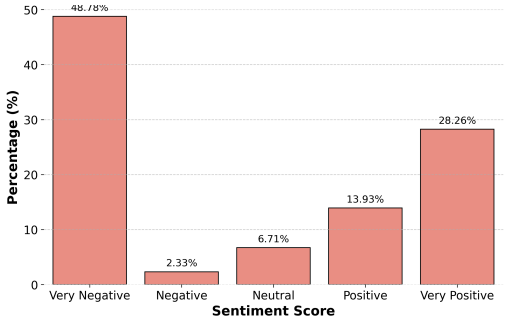


**Figure 4.4:** Result of Sentiment Analysis using BERT for COP26

Figures 4.3 and 4.4 shows the results gained from performing sentiment analysis using a a BERT model fine-tuned for sentiment analysis in multiple languages (multilingual). For this specific model, the sentiment classes are:1: Very negative 2:Negative 3: Neutral 4: Positive 5: Very positive. The chart shows that negative sentiments are the most frequent for both COP21 and COP26, with an increase in COP26. Neutral and positive sentiments are less frequent and follow a similar pattern, with COP26 showing lower frequencies than COP26. This may indicate the increasing concern about the topic.

Comparison of the results obtained from BERT on all data with the results when we omit coordination shows that excluding coordinated accounts does not have a significant impact on the sentiment distribution.

Furthermore, Figure4.5 and Figure4.6 give graphical representation of the distribution of sentiments among coordinated accounts and the comparison of the rates in COP21 and COP26. Based on these results, coordinated accounts, like the overall trend, exhibit more negative sentiments in COP26 compared to COP21. However, instead of simply shifting towards positive sentiments in COP26, there's a noticeable redistribution from the neutral category. This suggests a polarization or extremization of tweet tone, with coordinated accounts moving away from neutrality and adopting more pronounced positive or negative sentiments in COP26. An interesting observation is that, while some of their tweets in COP21 were identified as neutral, these sentiments in COP26 appear to have shifted mainly towards the positive, which is

against the overall trend of the sentiments.

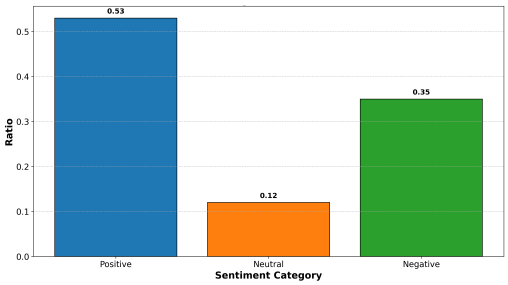


**Figure 4.5:** Ratio of Sentiment Categories for Coordinated accounts in COP21

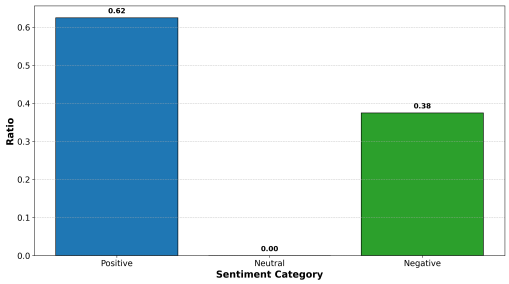


**Figure 4.6:** Ratio of Sentiment Categories for Coordinated accounts in COP26

### 4.2.2 Polarization Analysis:

Polarization in online discussions can be observed through the divergence of opinions into extreme viewpoints, leading to fragmented discourse. In our analysis of the COP21 and COP26 Twitter datasets, we aimed to measure the extent of this polarization by examining the sentiment distribution and ideological scoring and so we used the sentiment scores gained from our BERT model which was explained in the last section and we used network analysis described in chapter Materials & Methods for scoring the ideology. In addition, to go further with the study, we tried to find the correlation between the two. We conducted a rolling correlation analysis between the trend of sentiment scores and ideology scores to further investigate the relationship between them in the discourse around COP21 and COP26. In both COP21 and COP26 the correlation values fluctuate between positive and negative. This variability suggests that the influence of ideological positions on sentiment is not constant and therefore is not necessarily related.

The sentiment analysis for both COP21 and COP26 conferences revealed a significant presence of extreme sentiments. We observed a noticeable concentration of tweets expressing very negative or very positive emotions. This concentration suggests a polarized discourse, as the presence of extreme sentiments often correlates with a lack of common ground between opposing viewpoints.

We constructed interaction networks for both COP21 and COP26, incorporating data from retweets, mentions, and replies. For COP21, we created two separate networks—one based on retweets and another on mentions. These were then combined to form a comprehensive interaction graph. The same approach was applied to COP26. Node2Vec, a method used to generate numerical embeddings for nodes (e.g., users or tweets) based on their network relationships,

was utilized to analyze these networks. These embeddings were subjected to Principal Component Analysis (PCA) to compute ideology scores. The distribution of the results are shown in FIG4.7.

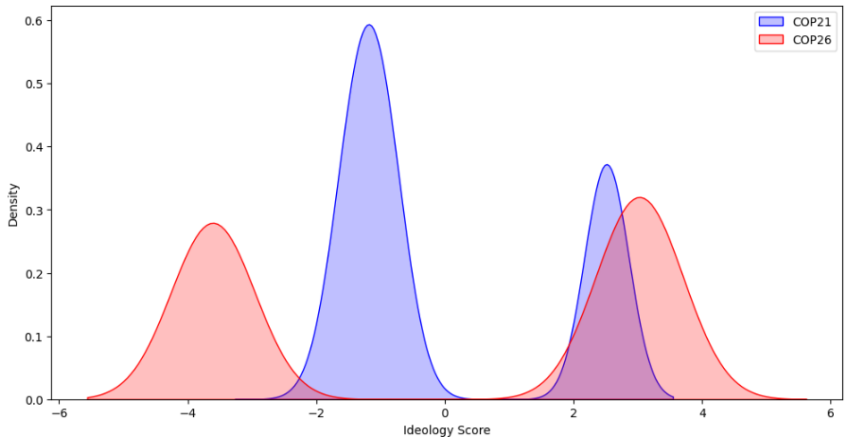


**Figure 4.7:** Distribution of Ideology Scores in COP21 and COP26

The broader range of ideology scores for COP26 suggests greater polarization or ideological diversity compared to COP21. The two peaks in the COP26 distribution are farther apart relative to those in COP21, indicating that communities are becoming more segregated and the debate more polarized. This results were also supported by the study done by Falkenberg et al.[10] that used a different method on the same data.

Hartigan's Dip Test[74] was applied to determine the modality of the ideology score distributions. This statistical test helps identify whether the distribution is bimodal, which would indicate the presence of distinct ideological groups—a common sign of polarization. Table4.3 indicates the results for our datasets, higher values suggest a stronger bimodal distribution and a p-value close to 0 suggests that the null hypothesis of a unimodal distribution can be rejected, indicating the presence of multimodality. The dip test results indicate that the ideological scores during COP21 exhibit a bimodal distribution, suggesting some degree of polarization. However, the p-value of 0.0 strongly supports this finding. In comparison, COP26 shows a higher dip test value than COP21, reflecting a more pronounced bimodal distribution. This indicates an increase in polarization, with the ideological landscape in COP26 becoming more distinctly divided into two separate groups.

In Table4.4 and Table4.5 we analyze 100 examples of random tweets in our datasets with the ideology scores of their authors and their sentiments. Additionally, we categorized the texts as either pro-climate or anti-climate based on their content and their support for the COP events.

| Event | Dip Test Value | P-Value |
|---|---|---|
| COP21 | 0.17 | 0.0 |
| COP26 | 0.23 | 0.0 |

**Table 4.3:** Hartigan's Dip Test Values and P-Values for COP21 and COP26

In the case of COP21, the majority of pro-climate texts exhibited moderate to high sentiment scores meaning in range of 1.2 to 3.5, indicating a generally positive public response to the initiatives discussed at the conference. The ideology scores were relatively moderate, suggesting that the discourse was inclusive of a range of perspectives, from cautious optimism to strong support for climate action. However, there were instances where texts had a high sentiment score but a lower ideology score, indicating support for the process but skepticism about its impact.

The analysis of COP26 data revealed a more polarized discourse compared to COP21. Although there were still many texts with high sentiment and ideology scores in range of 0.5 to almost 5, there was also a notable presence of critical voices, both from within the pro-climate community and from those opposing the event's outcomes. The sentiment was more mixed, with a significant number of texts expressing frustration or disappointment with the slow pace of progress. In addition, there were fewer texts with low ideology and high sentiment scores, indicating that the pro-climate discourse might be less accommodating of moderate or skeptical perspectives.

In general, we could categorize the ideology scores as low, moderate, and high. The low group tended to reflect skepticism, often questioning the potential effectiveness of the actions in COP21; In COP26 there was a noticeable increase in texts falling into the low ideology score category. The critiques in this group evolved to focus on the perceived lack of meaningful progress since COP21. The moderate ideology score category was well represented in COP21, reflecting a balanced perspective that acknowledged the importance of the Paris Agreement while calling for more robust commitments. However, the presence of moderate ideology scores decreased significantly during COP26, with fewer texts reflecting a neutral or cautiously optimistic stance. Moreover, in both COP21 and COP26 the high ideology score category was dominated by pro-climate advocates who fully support the events.

The correlation between ideology scores and sentiment scores was not consistent between both conferences. During both COP21 and COP26, high ideology scores often correlated with positive sentiment, especially among those strongly supporting climate initiatives. However, this correlation was not always straightforward. For example, some texts with high ide-

ology scores also carried a negative sentiment, reflecting frustration with the pace of progress or perceived shortcomings in the negotiations. Similarly, low ideology scores were typically associated with negative sentiment, but there were instances where low ideology scores carried a more neutral or mixed sentiment, especially when focusing on economic concerns rather than outright opposition to climate action.

| Text | Ideology Score | Sentiment Score | Climate Stance |
|---|---|---|---|
| Justifying a 'Humanitarian War' against Syria? The Sinister Role of the NGOs #COP21 | -1.19 | 1 | Anti-climate |
| @Reporterre COP21 il y a vraiment mieux à faire avec l'argent public que de l' engloutir dans un projet délétère | -1.95 | 1 | Anti-climate |
| RT WorldResources: Canada, EU, Japan, US: Post-2020 GHG Emissions Targets #COP21 | 1.50 | 1 | Neutral |
| RT DivestBerlin: We love those nasty honest advertising posters in Paris... #COP21 | 1.63 | 5 | Pro-climate |
| RT @Charlotteingrid: En skål för klimatavtalet! @mariaweimer och @KarinKarlsbro berättar om framgångarna med COP21 grönliberalism | 1.84 | 5 | Pro-climate |

**Table 4.4:** Examples from COP21 Data

| Text | Ideology Score | Sentiment Score | Climate Stance |
|---|---|---|---|
| The NZ Government's latest climate targets are woefully inadequate and fail to tackle dairy which means the rest of us pay the price. #COP26 | -2.37 | 1 | Anti-climate |
| RT @GMBScotOrg: BREAKING: Glasgow COP26 refuse and cleansing strike to go ahead. Glasgow City Council has acted in bad faith and failed to... #COP26 | 2.63 | 1 | Anti-climate |
| RT @JujuliaGrace: If you're embarrassed to see Boris Johnson not wearing a mask at COP26, please retweet. Let's send him a message. | -2.01 | 1 | Anti-climate |
| RT @Madhvi4EE: ClimateJustice is intergenerationaljustice. It's my future, children's future that is being negotiated | 2.34 | 5 | Pro-climate |
| Day 2 of our Climate Facts for Change. are sharing a new fact every day during COP26, to encourage you to take climate action. the Financial Times shared this chilling statistic. We must act now before it's too late. | 2.37 | 5 | Pro-climate |

**Table 4.5:** Examples from COP26 Data

In Figure4.8 we analyzed the ideology scores of coordinated accounts tweets. Given that the

distribution of ideology scores in COP26 has shifted further and shows higher density values, it indicates that the coordinated accounts have adopted more consistent opinions over time.

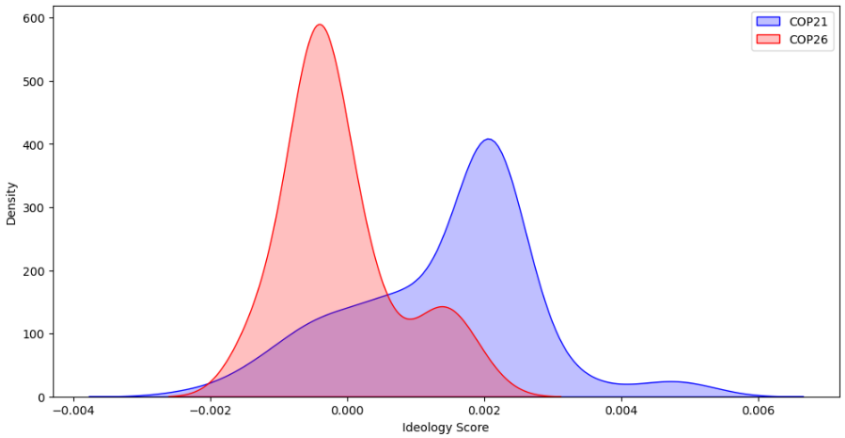


**Figure 4.8:** Distribution of Ideology Scores of Coordinated Accounts in COP21 and COP26

## 4.3 Toxicity & Skepticism Analysis

### 4.3.1 Toxicity Levels

We examined the levels of toxicity within our datasets related to COP21 and COP26 and specifically compared toxicity levels between coordinated and non-coordinated accounts. we employed two complementary approaches using advanced tools in both R and Python.

First we leveraged the Perspective API in R, a tool developed by Jigsaw and Google[75]. This API provides detailed scores across various dimensions of toxicity, including TOXICITY and SEVERE TOXICITY. The results are shown in Figure4.9 which indicate that the average toxicity score for COP26 is higher than that for COP21, indicating greater overall toxicity in COP26. Similarly, the average severe toxicity score for COP26 (0.00363) is also higher than that for COP21 (0.00271). The plot uses light blue bars for COP21 and red bars for COP26, with side-by-side comparisons for each metric, visually highlighting the differences in toxicity levels between the two datasets.

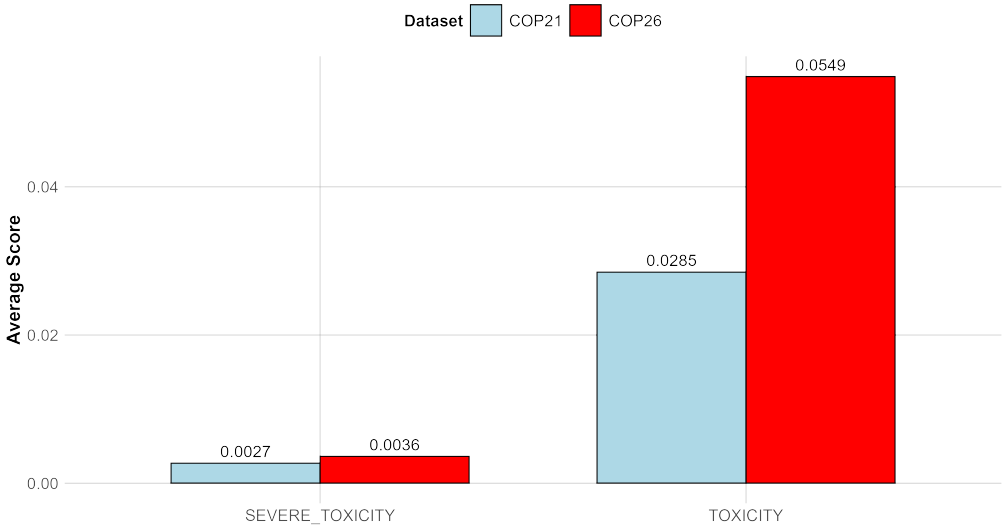


**Figure 4.9:** Toxicity and Severe Toxicity Rates in COP21 and COP26 resulted from Perspective API

Moving forward, we analyzed the toxicity of coordinated accounts and as the Table4.6 indicates, the toxicity in coordinated accounts tweets were relatively lower than average. This suggests that coordinated accounts may be more strategic in their language use, potentially to avoid detection or to maintain credibility in the discourse.

| Event | Average Toxicity | Average Severe Toxicity |
|---|---|---|
| COP21 | 0.027 | 0.0021 |
| COP26 | 0.050 | 0.0031 |

**Table 4.6:** Toxicity of Coordinated Accounts in COP21 and COP26 resulted from Perspective API

To have a more thorough analysis, we also used the Unitary/toxic-bert model, a BERT model fine-tuned specifically for detecting toxic language. The model assigns a toxicity score to each tweet, where higher scores indicate a greater level of toxicity. In Figure4.10 and Figure4.11 we can see the distribution of toxicity in COP21 and COP26 respectively. A toxicity score closer to 1 signifies more toxic language. Similarly to our findings with the Perspective API, but not as extreme, these figures reveal a noticeable increase in the use of toxic language during COP26, highlighting a shift toward more hostile and inflammatory discourse over time.

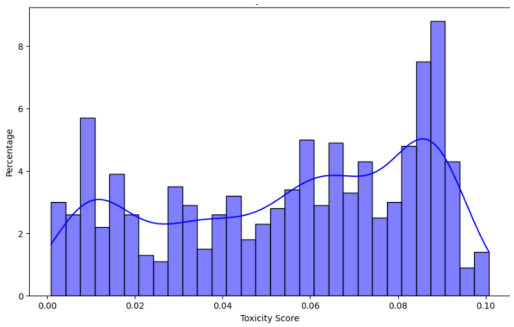


**Figure 4.10:** Result of Toxicity Analysis using BERT for COP21

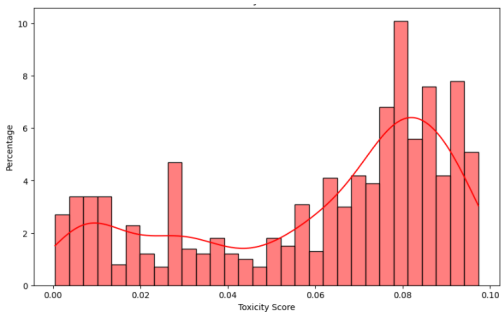


**Figure 4.11:** Result of Toxicity Analysis using BERT for COP26

Moreover, we analyzed the toxcity levels around the dates of the two conferences using BERT and the results are shown in Figure4.12 and Figure4.13. This analysis shows that on average, around the time of the COP21 conference, the levels of toxicity were higher in the discussions despite the higher overall toxicity levels in COP26.

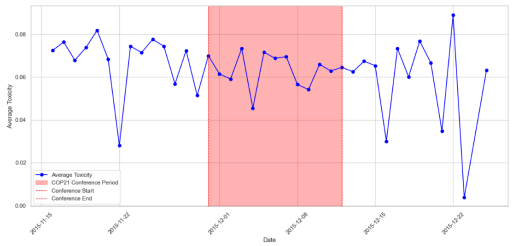


**Figure 4.12:** Analysis of Daily Toxicity Levels using BERT for COP21

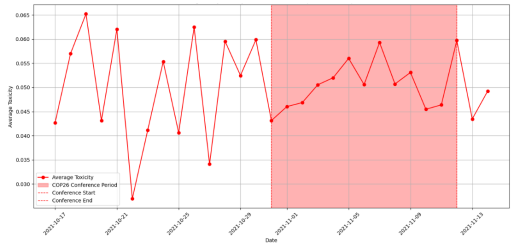


**Figure 4.13:** Analysis of Daily Toxicity Levels using BERT for COP26

Similarly to the findings from the Perspective API, the BERT analysis also reveals that the toxicity levels of coordinated accounts are not higher than the average.

Comparison of tweets with positive and negative sentiments indicated a significant difference in their toxicity levels.

### 4.3.2 Skepticism

Finally, we implemented another dual approach to identify tweets that expressed doubt or disbelief in climate science or the urgency of climate action. This analysis focused on detecting skepticism and examining its relationship with toxicity in the discourse.

The first method involved identifying skeptical tweets by searching for specific keywords commonly associated with climate skepticism. These keywords included terms like ”hoax,”

”scam,” ”climate alarmism,” ”fake,” ”fraud,” ”lies,” ”myth,” ”swindle,” ”nonsense,” and ”conspiracy.” Table4.7 shows the percentages of skeptic tweets resulted from this method which indicates the tweets were slightly more skeptical during COP26 compared to COP21.

| Event | Skeptics Rate |
|---|---|
| COP21 | 1.43% |
| COP26 | 1.78% |

**Table 4.7:** Skeptic Tweets based on Keyword-Based Analysis

To complement the keyword-based approach and enhance the accuracy of skepticism detection with a better approach, we employed sentiment classification using a BERT model ”monologg/bert-base-cased-goemotions-original”. This model classified the emotional tone of tweets into categories such as ”anger,” ”fear,” and ”surprise.” We focused particularly on the categories of ”confusion”, ”fear”, and ”nervousness”, which are often indicative of skeptical or oppositional sentiments toward climate change. This method allowed us to capture more nuanced expressions of skepticism, including those that might not use explicit keywords but still conveyed doubt or resistance. Figure4.14 and Figure4.15 illustrates the results gained from this method which show the same increasing trend.

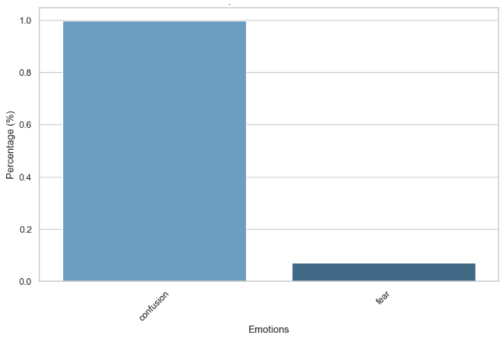


**Figure 4.14:** Analysis of Emotions Connected to Skepticism in COP21

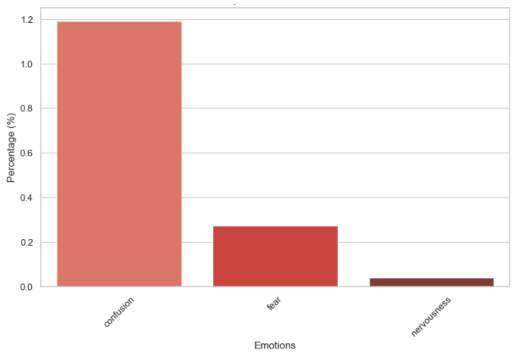


**Figure 4.15:** Analysis of Emotions Connected to Skepticism in COP26

Analyzing skepticism in tweets from coordinated accounts, as illustrated in Table4.8, reveals that even though the skepticism rates are lower in coordinated accounts in general, a higher proportion of these tweets express skeptical concerns in COP26. Moreover, they also show more confusion and fear. This suggests that coordinated campaigns may be using skepticism as a strategic tool to influence public opinion. This trend emphasizes the need for further exami-

nation of how coordinated efforts exploit skepticism to shape and polarize public discourse on climate change.

| Event | Coordinated Skeptic Tweets |
|---|---|
| COP21 | 0.06% |
| COP26 | 1.09% |

**Table 4.8:** Skeptic Tweets based on Keyword-Based Analysis

Moreover, Figure4.16 indicates the distribution of ideology scores in skeptic tweets. It shows that while skepticism during COP21 was largely concentrated on the lower side of the ideological spectrum, in COP26 skeptic tweets spread more evenly across a broader range of ideological groups, with views distributed more equally between the two sides.

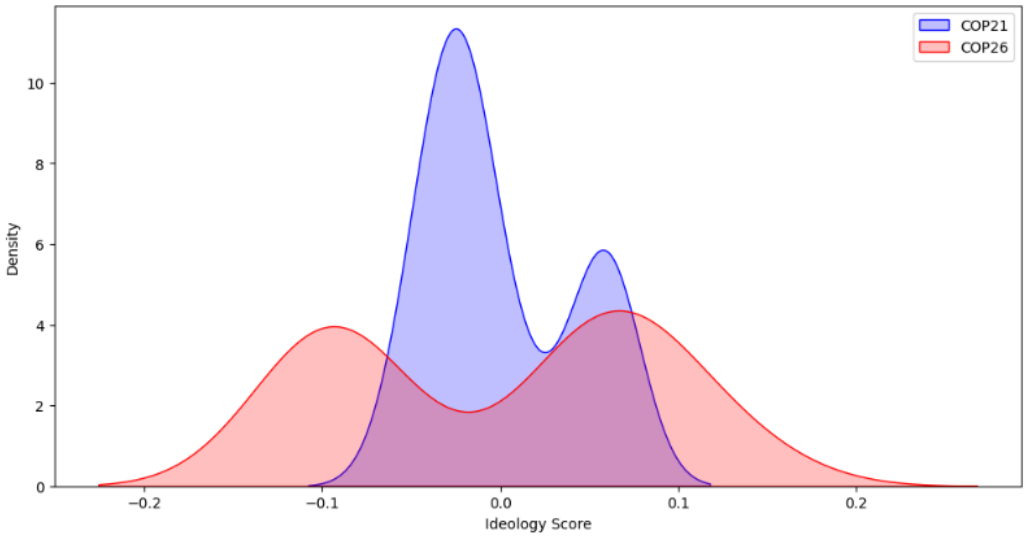


**Figure 4.16:** Ideology Score Distribution of Skeptic Tweets in COP21 and COP26

# 5 Related Work

## 5.1 Present Works

In this thesis we explored the dynamics of climate change discussions on social media and a comprehensive examination of coordination, polarization, and toxicity. This chapter reviews key studies and methodologies related to these factors, highlighting their role in shaping online discourse. By focusing on how these elements interact during significant climate discussion events like COP21 and COP26, we provide a contextual foundation for analyzing the trends and correlations observed in our study in the next chapter.

Social media has increasingly become pivotal in shaping public perception and discourse surrounding global warming and climate change events. Twitter plays a crucial role in these discussions, primarily due to its widespread use by politicians, public figures, and organizations. The platform offers a space where influential individuals actively engage with their audiences, share their views, and shape public discourse. Politicians and public figures use Twitter to broadcast their messages, mobilize support, and influence public opinion in real-time. This high level of engagement from key stakeholders amplifies the impact of discussions and controversies surrounding events like COP21 and COP26, making Twitter a central arena for understanding diverse and often subjective viewpoints. Research indicates that discussions on global warming on Twitter tend to be highly subjective, with a subjectivity score of 62.46% [76]. This subjectivity can contribute to the spread of misinformation and affect the public understanding of

climate issues. Furthermore, studies have shown that social media content on climate change often highlights local events that might be underrepresented in mainstream media[77]. This allows for a more comprehensive view of public sentiment and emerging trends related to climate action and policy.

In this context, analyzing social media data from major climate change events, such as COP21 and COP26, provides valuable insights into how these platforms contribute to the discourse on climate change. By examining the social media data, we can better understand the dynamics of online discussions, including the prevalence of subjective content[78], the impact of events[10], and the general sentiment and polarization of climate change debates[10][70]. Given that climate change is one of the most pressing challenges facing our society, understanding the dynamics of public discourse around major climate events such as COP21 and COP26 is of paramount importance. Climate change impacts numerous aspects of our lives, including environmental sustainability, economic stability, and public health. As such, studying the debate on climate change is crucial to ensure informed and constructive discussions. This examination helps to mitigate risks such as excessive polarization, misinformation, and the potential for reduced public engagement. By analyzing how different stakeholders communicate and interact on platforms like Twitter, we can better understand the factors influencing public perception and policy support, ultimately contributing to more effective climate action and resilience strategies.

Recent research has made substantial progress in identifying and analyzing coordinated behavior on social media platforms. A range of methods has been developed to detect and understand coordination, each offering different levels of complexity and applicability. For example, Yu et al.[79] proposed a framework that detects coordination by analyzing shared media content within short time intervals, focusing on the 2019 Philippine elections. This approach highlights how timely content sharing can signal coordinated activities.

Sharma et al.[29] introduced a generative model that combines temporal point processes with Gaussian mixture models to capture influence patterns and detect anomalous group behaviors. This method provides a nuanced understanding of how coordinated groups behave and influence online discussions.

Further advancements are seen in the work by Nizzoli et al.[23], who developed a network-based approach to uncover various coordination patterns and estimate coordination degrees within communities, particularly during the 2019 UK General Election. This approach emphasizes the importance of analyzing network structures to reveal how coordination manifests within different online communities.

The Coordination Network Toolkit, presented by Graham et al.[35], represents a significant development in this field. This open-source package enables multi-behavior coordination detection across platforms by using weighted, directed multigraphs. The toolkit facilitates the identification of complex coordination patterns and provides valuable information on online influence, disinformation campaigns, and digital activism.

These studies collectively emphasize the importance of analyzing coordinated behaviors to understand its influence on the dissemination of online information. The proposed methods range from simple content-sharing metrics to sophisticated statistical models, each offering unique insights into how coordination can impact social media discussions. In the context of climate change discussions during COP21 and COP26, applying these methods will enhance our ability to detect and analyze coordinated activities.

Moving forward, sentiment analysis has become a crucial tool in social media research, driven by its wide-ranging applications across various domains, including public health, politics, and marketing. The fundamental goal of sentiment analysis is to computationally process and understand the opinions and emotions expressed in user-generated content, such as text, images, and videos shared on platforms like Twitter and Facebook. By doing so, researchers can measure public sentiment on specific topics, monitor changes in public opinion, and even predict future trends based on the emotional tone of online discussions.

In recent years, BERT has emerged as a powerful tool for sentiment analysis, especially in analyzing large-scale social media data. BERT's transformer-based architecture allows it to capture the contextual nuances of language, making it particularly effective in understanding the often informal and complex language used on social media platforms[80]. This capability has made BERT especially useful during the COVID-19 pandemic, where it was used to analyze public sentiment, providing critical insights into public concerns and attitudes[81].

Studies have demonstrated that BERT outperforms traditional machine learning algorithms in sentiment classification tasks, achieving high accuracy levels on large datasets. For instance, Thulasi Bikku et al.[82] reported that BERT-based models achieved superior performance compared to older methods, particularly in tasks involving complex and nuanced sentiment classification. This is largely due to BERT's ability to process contextual information more effectively than models like Naive Bayes, which typically rely on simpler, bag-of-words approaches [83] [84].

During the COVID-19 pandemic, researchers applied BERT to analyze social media discussions and were able to achieve validation accuracies of around 94% in sentiment classification tasks[85]. A specific study on the Chinese social media platform Sina Weibo used BERT to

classify sentiments and identify key public concerns related to the pandemic, such as the origin of the virus, symptoms, production activities, and public health control measures [86]. This study exemplifies how BERT's ability to handle context and language nuances can be crucial for understanding public sentiment during crises. Studies have applied multilingual BERT models to various languages, including Bahasa Indonesia for movie reviews[87].

Moreover, BERT's flexibility and adaptability have been further demonstrated through comparative analyses of different variants, such as RoBERTuito and RuBERT. These studies have shown that models like RoBERTuito achieve the highest accuracy, at 83.23%, when fine-tuned for specific languages and contexts[88]. Additionally, BERT has proven adept at managing Twitter-specific linguistic challenges, including slang and sarcasm, which are common on social media [84]. Despite its strengths, researchers have identified certain limitations of BERT, such as its dependency on large datasets and computational resources, and have suggested future directions like the integration of external knowledge to enhance its performance[84].

Overall, on online platform like Twitter BERT-based model demonstrated superior performances with respect to baseline methods in sentiment analysis. This success underscores BERT's importance in modern sentiment analysis, making it a key tool for researchers aiming to understand public sentiment in social media contexts[84][83].

Sentiment analysis, is also used to detect and identify opinions in text, which can be applied to measure polarization in social media by analyzing the sentiment of user comments[7]. A study by Alsinet et al.[89] created a measure specifically for online debates, combining user agreement partitioning with sentiment analysis of inter-group interactions. Their study on Reddit discussions demonstrated the measure's ability to capture varying levels of polarization across different topics. Another study by Singh et al.[90] presented a multi-opinion based method that extends beyond binary scenarios, testing it on various networks and a Twitter case study about COVID-19 treatments.

Furthermore, a study by Belcastro et al. [91] introduced the IOM-NN methodology, which utilizes neural networks to assess user polarization during elections, achieving high accuracy compared to traditional sentiment analysis methods. Studies have analyzed polarization through various lenses, including ideological stances, behavioral interactions, and network structures.

Network analysis offers valuable tools for identifying and measuring polarization in various contexts. Researchers have proposed different approaches to quantify polarization, including node homophily distributions[92] and generalized axiomatic characterizations for node- and edge-weighted networks[93]. Methods for detecting polarization encompass techniques based on homophily, modularity, random walks, and balance theory[94].

Furthermore, The rise of offensive and inflammatory content on platforms like Twitter has led to the development of advanced computational models aimed at identifying and reducing toxic behavior[95]. A central approach in this field involves the use of BERT-based models, which have been proven to be highly effective in this context.

Recent studies have demonstrated the versatility and power of BERT in detecting toxic language. For example, a study by Karimi et al. [96] introduced a model that combines CharacterBERT [97] with a traditional bag-of-words model to detect toxic spans within text. This hybrid approach achieved competitive results, highlighting the effectiveness of integrating character-level and word-level features in toxicity detection. Similarly, another study by Yoshida et al. [98] proposed a BERT-based system designed not only to predict potentially inflammatory posts but also to transform toxic expressions into safer alternatives. This innovative approach underscores the potential of BERT in not just identifying but also actively mitigating toxic language.

Another significant contribution to this field is the work by Luu and Nguyen[99], who employed a BiLSTM-CRF model enhanced with ToxicBERT for toxic span detection. Their model demonstrated high efficacy of combining BERT with traditional sequence labling techniques for precise toxicity detection. Further advancements were made by Singh et al. [100], who developed a toxic comment analyzer using BERT, achieving 97% accuracy on benchmark datasets. These studies collectively highlight the robust capabilities of BERT-based models in addressing various aspects of online toxicity, from character-level analysis to sentence-level classification and even text modification.

In addition to academic research, practical tools like the Perspective API have gained traction in combating online toxicity[52]. The Perspective API, which applies machine learning to evaluate the toxicity of comments, offers a scalable solution for platforms looking to filter and manage user-generated content. Although specific studies on the Perspective API are limited, its application in real-world settings demonstrates the growing importance of automated tools in maintaining healthy online environments.

Skepticism in political social media data is influenced by several factors, including the prevalence of disinformation, the role of algorithms, and user behavior. A study showed that significant portion of Turkish students distrust political news on social media, with 57.4% not following leaders of their preferred parties and 56.4% not following leaders of opposing parties, largely due to perceived disinformation and lack of credibility[101]. This distrust is compounded by the fact that political content constitutes a small fraction of social media feeds, making it challenging to gauge its real-world impact[78]. Skepticism also leads to passive consumption of

news and it is reflected in the reluctance of people to engage with political content, as they often use social media for other interests like entertainment and sports[101]. The role of algorithms in creating echo chambers further exacerbates this issue, as users are often exposed to content that aligns with their preexisting beliefs, limiting their exposure to diverse viewpoints[101].

Despite these challenges, social media can still play a role in political engagement and depolarization. Some users find common ground with opposing party supporters, suggesting that social media can facilitate civil discourse if used thoughtfully. However, the overall effect of social media on political polarization is considered low to moderate, influenced by the lack of trust in media and politicians and the alternative uses of social media[101].[102]

## 5.2 Identification of Gaps in the Literature

While significant advancements have been achieved in understanding the dynamics of social media discussions, particularly around climate change, there are still critical gaps that need to be addressed. One major gap is the lack of a comprehensive, cross-event analysis that spans multiple COP conferences. Most existing studies tend to focus on isolated events or short time frames, failing to capture the broader trends and shifts in discourse that occur over longer periods or across different global contexts. Furthermore, the correlations between coordinated behavior, polarization, sentiment, and toxicity within these discussions are not well understood.

This thesis aims to bridge these gaps by conducting a detailed examination of coordinated behavior and its subsequent impact on sentiment and toxicity in climate change discussions on social media. By applying advanced and up-to-date techniques across multiple COP events, this study not only seeks to validate the effectiveness of existing tools but also to reveal new insights into how coordinated efforts shape public discourse on climate change. This approach will provide a more nuanced understanding of the dynamics at play, offering valuable contributions to both academic research and practical applications in policy-making and public communication strategies.

# 6
# Conclusion

This thesis addresses the complex dynamics of online discourse surrounding climate change, with a particular focus on the role of coordinated accounts. As online platforms become increasingly impactful in shaping public opinion, understanding how these accounts manipulate discourse is crucial. The problem at hand involves analyzing how coordinated efforts impact the polarization, toxicity, and sentiment of discussions related to major climate conferences, such as COP21 and COP26. We also aimed to analyze the relationship between coordination and polarization, assess the correlation between coordination on toxicity levels, and analyze the evolution of coordination, toxicity, and polarization between COP21 and COP26.

We performed a comprehensive analysis of online discussions, utilizing the latest methods and cross-examining their results to gain a thorough understanding. We used behavioral and network analysis to detect coordinated behavior, and further evaluated them with the CooRTweet R package.

For sentiment analysis, we integrated the Syuzhet R library and BERT models to capture emotional tones in tweets. We assessed polarization through sentiment distribution and ideological scoring, supported by network analysis and PCA. Additionally, we explored trends over time to understand how discourse evolved.

Skepticism was analyzed using keyword-based methods and sentiment classification with BERT and Toxicity levels were evaluated with the Perspective API and the toxic BERT model, offering valuable insights into different dimensions of toxicity.

This multifaceted approach provided a robust evaluation and deep insights into the dynam-

ics of online interactions.

The results of this research indicate an increase in the rate of coordination from COP21 to COP26, that could possibly signal the increased effort to influence online debates. Furthermore, our analysis revealed shifts in sentiment, polarization, and toxicity over time, highlighting evolving dynamics in online discussions.

The ideology scores ranged from low (indicating skepticism) to moderate (reflecting balanced views) to high (showing strong support). The ideological landscape became clearly more polarized during COP26 which would explain the higher levels of toxicty among positive sentiment, that is a sign of highly polarized discussions; people may use strong, positive affirmations to support their side while simultaneously attacking the other side in the same message. Meanwhile, negative sentiments increased overall, while positive and neutral sentiments decreased.

The correlation between sentiment and ideology was inconsistent. High ideology scores usually matched positive sentiment, but some texts expressed frustration despite strong support. Low ideology scores often correlated with negative sentiment, though some were more neutral, focusing on economic issues rather than outright opposition.

Moreover, toxicity levels were notably higher during COP26. However, interestingly, coordinated accounts exhibited relatively lower levels of toxicity, suggesting a strategic use of language to maintain credibility, since toxicity on social media significantly impacts public engagement and often leads to withdrawal from conversations and increase in social distance between groups[50].

Furthermore, skepticism seemed to be increasing in COP26 compared to COP21 as well. This might reflect growing doubts or criticisms regarding climate change policies, the effectiveness of conferences, or the general handling of climate issues. This result was supported by analyzing the ideology scores of skeptical tweets that indicated the same trend in growing skepticism in a broader range of ideology.

Based on these results, there is a growing attempt of coordinated accounts with certain behaviors that influence online discourse. Sentiment analysis reveals that coordinated accounts mostly convey positive sentiments, suggesting a deliberate effort to promote favorable narratives. Furthermore, analyzing their ideology score shows a shift towards more negative scores in the same range. This followed the overall trend of ideology in COP26 compared to COP21. However, despite these efforts, they appear to be less influential than expected in information cascades, potentially due to inefficient placement strategies and limited resources.

Compared to the existing literature, there are several advantages to our study. Firstly, we combined metrics for coordination identification to provide a more robust and comprehensive

approach and validated them to enhance the accuracy and reliability of coordination detection. Secondly, for sentiment and polarization analyses, we used the state-of-the-art approach to have a more nuanced understanding of the emotional tone of tweets and capture a wide range of sentiments and ideologies. Thirdly, the use of the Perspective API and the BERT model for toxicity and skepticism analysis offered a very detailed assessment of the matter. So, overall one of the strengths of this thesis was using the state-of-the-art approach and integrating them with other approaches. In addition, we enhanced our understanding of the influence of coordinated behavior on the evolution of online debates by comparing data from COP21 and COP26.

In conclusion, this thesis has provided a comprehensive analysis of the dynamics of coordinated behavior, sentiment, polarization, toxicity, and skepticism in online debates surrounding climate change. By employing a multifaceted approach, we have advanced the understanding of coordinated accounts and their impact on public discourse. The findings highlight the importance of continuing research and the development of strategies to mitigate the impact of coordinated behavior on online debates. As we navigate the challenges of misinformation and polarization in the digital age, fostering a more inclusive and informed public discourse is crucial to address issues such as climate change.

# References


[1] D. Miller, J. Sinanan, X. Wang, T. McDonald, N. Haynes, E. Costa, J. Spyer, S. Venkatraman, and R. Nicolescu, *How the world changed social media*. UCL press, 2016.

[2] M. Cinelli, S. Cresci, W. Quattrociocchi, M. Tesconi, and P. Zola, "Coordinated inauthentic behavior and information spreading on twitter," *Decision Support Systems*, vol. 160, p. 113819, 2022.

[3] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media: methods and case studies," in *Proceedings of the international AAAI conference on web and social media*, vol. 15, 2021, pp. 455–466.

[4] T. Reuter, "Achieving global justice, security and sustainability: Compassion as a transformative method," *Cadmus*, vol. 4, no. 5, pp. 30–37, 2021.

[5] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, and P. Nakov, "The spread of propaganda by coordinated communities on social media," in *Proceedings of the 14th ACM Web Science Conference 2022*, 2022, pp. 191–201.

[6] D. Weber and F. Neumann, "Amplifying influence through coordinated behaviour in social networks," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 111, 2021.

[7] V. Maslej-Krešňáková, M. Sarnovskỳ, P. Butka, and K. Machová, "Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification," *Applied Sciences*, vol. 10, no. 23, p. 8631, 2020.

[8] J. Zarocostas, "How to fight an infodemic," *The lancet*, vol. 395, no. 10225, p. 676, 2020.

[9] D. Caled and M. J. Silva, "Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation," *Journal of Computational Social Science*, vol. 5, no. 1, pp. 123–159, 2022.

[10] M. Falkenberg, A. Galeazzi, M. Torricelli, N. Di Marco, F. Larosa, M. Sas, A. Mekacher, W. Pearce, F. Zollo, W. Quattrociocchi *et al.*, "Growing polarization around climate change on social media," *Nature Climate Change*, vol. 12, no. 12, pp. 1114–1121, 2022.

[11] F. Barbero, S. op den Camp, K. van Kuijk, C. Soto García-Delgado, G. Spanakis, and A. Iamnitchi, "Multi-modal embeddings for isolating cross-platform coordinated information campaigns on social media," in *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer, 2023, pp. 14–28.

[12] B. Van De Velde, A. Meijer, and V. Homburg, "Police message diffusion on twitter: analysing the reach of social media communications," *Behaviour & information technology*, vol. 34, no. 1, pp. 4–16, 2015.

[13] T. Magelinski, L. Ng, and K. Carley, "A synchronized action framework for responsible detection of coordination on social media, 2021," *CoRR. arXiv: abs/2105.07454*.

[14] Statista. (2024) Number of social media users worldwide from 2017 to 2028 (in billions). [Online; accessed May 17, 2024]. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[15] ——. (2024) Daily time spent on social networking by internet users worldwide from 2012 to 2024 (in minutes). [Online; accessed February 22, 2024]. [Online]. Available: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

[16] C. T. Carr and R. A. Hayes, "Social media: Defining, developing, and divining," *Atlantic journal of communication*, vol. 23, no. 1, pp. 46–65, 2015.

[17] J. J. Van Bavel, C. E. Robertson, K. Del Rosario, J. Rasmussen, and S. Rathje, "Social media and morality," *Annual review of psychology*, vol. 75, no. 1, pp. 311–340, 2024.

[18] M. F. Schober, J. Pasek, L. Guggenheim, C. Lampe, and F. G. Conrad, "Social Media Analyses for Social Measurement," *Public Opinion Quarterly*, vol. 80, no. 1, pp. 180–211, 01 2016. [Online]. Available: https://doi.org/10.1093/poq/nfv048

[19] R. Kitchin, "Big data, new epistemologies and paradigm shifts," *Big Data & Society*, vol. 1, no. 1, p. 2053951714528481, 2014. [Online]. Available: https://doi.org/10.1177/2053951714528481

[20] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan, and S. Liu, "Fake news on social media: the impact on society," *Information Systems Frontiers*, vol. 26, no. 2, pp. 443–458, 2024.

[21] S. DAS and R. PRAKASH, "Fake news evaluation awareness level on social media in guwahati: A survey."

[22] P. B. J. S. A. F. S. Eileen Culloty, Padraig Murphy and D. Zhang, "Researching visual representations of climate change," *Environmental Communication*, vol. 13, no. 2, pp. 179–191, 2019. [Online]. Available: https://doi.org/10.1080/17524032.2018.1533877

[23] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, and M. Tesconi, "Coordinated behavior on social media in 2019 uk general election," in *Proceedings of the international AAAI conference on web and social media*, vol. 15, 2021, pp. 443–454.

[24] T. Magelinski, L. H. X. Ng, and K. M. Carley, "A synchronized action framework for responsible detection of coordination on social media," 2021. [Online]. Available: https://arxiv.org/abs/2105.07454

[25] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on twitter: How to coordinate a disinformation campaign," *Political communication*, vol. 37, no. 2, pp. 256–280, 2020.

[26] D. Schoch, F. B. Keller, S. Stier, and J. Yang, "Coordination patterns reveal online political astroturfing across the world," *Scientific reports*, vol. 12, no. 1, p. 4572, 2022.

[27] F. Giglietto, N. Righetti, L. Rossi, and G. Marino, "It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 italian elections," *Information, Communication & Society*, vol. 23, no. 6, pp. 867–891, 2020.

[28] K. Sharma, E. Ferrara, and Y. Liu, "Identifying coordinated accounts in disinformation campaigns," 2020.

[29] K. Sharma, Y. Zhang, E. Ferrara, and Y. Liu, "Identifying coordinated accounts on social media through hidden influence and group behaviours," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1441–1451.

[30] S. Tardelli, L. Nizzoli, M. Avvenuti, S. Cresci, and M. Tesconi, "Multifaceted online coordinated behavior in the 2020 us presidential election," *EPJ Data Science*, vol. 13, no. 1, p. 33, 2024.

[31] A. Kretser, D. Murphy, S. Bertuzzi, T. Abraham, D. B. Allison, K. J. Boor, J. Dwyer, A. Grantham, L. J. Harris, R. Hollander *et al.*, "Scientific integrity principles and best practices: recommendations from a scientific integrity consortium," *Science and Engineering Ethics*, vol. 25, pp. 327–355, 2019.

[32] L. Mannocci, M. Mazza, A. Monreale, M. Tesconi, and S. Cresci, "Detection and characterization of coordinated online behavior: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2408.01257

[33] B. De Clerck, J. C. F. Toledano, F. Van Utterbeeck, and L. E. Rocha, "Detecting coordinated and bot-like behavior in twitter: the jürgen conings case," *EPJ Data Science*, vol. 13, no. 1, p. 40, 2024.

[34] M. Jurišić, I. Tomičić, and P. Grd, "User behavior analysis for detecting compromised user accounts: A review paper," *Cybernetics and Information Technologies*, vol. 23, no. 3, pp. 102–113, 2023.

[35] T. Graham, S. Hames, and E. Alpert, "The coordination network toolkit: a framework for detecting and analysing coordinated behaviour on social media," *Journal of Computational Social Science*, pp. 1–22, 2024.

[36] Y. Zhang, K. Sharma, and Y. Liu, "Capturing cross-platform interaction for identifying coordinated accounts of misinformation campaigns," in *European Conference on Information Retrieval*. Springer, 2023, pp. 694–702.

[37] C. R. Sunstein, "Deliberative trouble? why groups go to extremes," in *Multi-party dispute resolution, democracy and decision-making*. Routledge, 2017, pp. 65–95.

[38] S. Liu and H. Wen, "Agent-based modelling of polarized news and opinion dynamics in social networks: a guidance-oriented approach," *Journal of Complex Networks*, vol. 12, no. 4, p. cnae028, 07 2024. [Online]. Available: https://doi.org/10.1093/comnet/cnae028

[39] L. T. L. Terren and R. B.-B. R. Borge-Bravo, "Echo chambers on social media: A systematic review of the literature," *Review of Communication Research*, vol. 9, 2021.

[40] M. Nordbrandt, "Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties," *New Media Society*, vol. 25, pp. 3392 – 3411, 2021.

[41] T. Jiang, "Studying opinion polarization on social media," *Social Work and Social Welfare*, vol. 4, no. 2, pp. 232–241, 2022.

[42] E. Kubin and C. von Sikorski, "The role of (social) media in political polarization: a systematic review," *Annals of the International Communication Association*, vol. 45, no. 3, pp. 188–206, 2021. [Online]. Available: https://doi.org/10.1080/23808985.2021.1976070

[43] C. Treuillier, S. Castagnos, and A. Brun, "A multi-factorial analysis of polarization on social media," 2023. [Online]. Available: https://arxiv.org/abs/2306.00032

[44] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," 2018. [Online]. Available: https://arxiv.org/abs/1802.01400

[45] D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," 2019. [Online]. Available: https://arxiv.org/abs/1904.01596

[46] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[47] H. Kim, "Sentiment analysis: Limits and progress of the syuzhet package and its lexicons." *DHQ: Digital Humanities Quarterly*, vol. 16, no. 2, 2022.

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

[49] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[50] J. D. Gallacher, M. W. Heerdink, and M. Hewstone, "Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters," *Social Media + Society*, vol. 7, 2021.

[51] A. Sheth, V. L. Shalin, and U. Kursuncu, "Defining and detecting toxicity on social media: context and knowledge are key," *Neurocomputing*, vol. 490, pp. 312–318, 2022.

[52] J. Salminen, S. Sengün, J. Corporan, S.-g. Jung, and B. J. Jansen, "Topic-driven toxicity: Exploring the relationship between online toxicity and news topics," *PloS one*, vol. 15, no. 2, p. e0228723, 2020.

[53] M. Avalle, N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli *et al.*, "Persistent interaction patterns across social media platforms and over time," *Nature*, vol. 628, no. 8008, pp. 582–589, 2024.

[54] J. W. Kim, A. Guess, B. Nyhan, and J. Reifler, "The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity," *Journal of Communication*, vol. 71, no. 6, pp. 922–946, 2021.

[55] D. Noever, "Machine learning suites for online toxicity detection," *arXiv preprint arXiv:1810.01869*, 2018.

[56] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos, "Toxicity detection: Does context really matter?" *arXiv preprint arXiv:2006.00998*, 2020.

[57] A. Alsharef, K. Aggarwal, Sonia, D. Koundal, H. Alyami, and D. Ameyed, "[retracted] an automated toxicity classification on social media using lstm and word embedding," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 8467349, 2022.

[58] Y. Chakraverty, A. Kaintura, B. Kumar, A. Khanna, M. Sharma, and P. K. Pareek, "Analyzing the feasibility of bert model for toxicity analysis of text," in *International Conference on Innovative Computing and Communications*, A. E. Hassanien, O. Castillo, S. Anand, and A. Jaiswal, Eds. Singapore: Springer Nature Singapore, 2024, pp. 653–661.

[59] A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, "A new generation of perspective api: Efficient multilingual character-level transformers," 2022. [Online]. Available: https://arxiv.org/abs/2202.11176

[60] E. Merkley and D. A. Stecula, "Party cues in the news: Democratic elites, republican backlash, and the dynamics of climate skepticism," *British Journal of Political Science*, vol. 51, no. 4, pp. 1439–1456, 2021.

[61] J. L. Tavani, A. Piermattéo, G. Lo Monaco, and S. Delouvée, "Skepticism and defiance: Assessing credibility and representations of science," *PloS one*, vol. 16, no. 9, p. e0250823, 2021.

[62] Ž. Pavića and E. Kovačevićb, "Negative information leads to a decline of trust in science: the connection between traditional and social media uses and vaccination conspiracy beliefs," *Journal of Community Positive Practices*, no. 2, pp. 51–77, 2024.

[63] S. Altay, "How effective are interventions against misinformation?" 2022.

[64] C. Boussalis and T. G. Coan, "Signals of doubt: Text-mining climate skepticism," in *Workshop", London School of Economics*, vol. 27. Citeseer, 2013, p. 28.

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

[66] N. Righetti and P. Balluff, "CooRTweet: An R package to detect coordinated networks on Twitter," May 2023. [Online]. Available: https://github.com/nicolarighetti/CooRTweet

[67] J. Jockers, *syuzhet: Extract Sentiment and Plot Sentiment Profiles*, 2024, r package version 1.0.7. [Online]. Available: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

[68] N. Town, "Bert base multilingual uncased sentiment model," https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment, 2024.

[69] B. Savani, "Distilbert base uncased emotion classification model," https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion, 2024.

[70] J. Flamino, A. Galeazzi, S. Feldman, M. W. Macy, B. Cross, Z. Zhou, M. Serafino, A. Bovet, H. A. Makse, and B. K. Szymanski, "Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections," *Nature Human Behaviour*, vol. 7, no. 6, pp. 904–916, 2023.

[71] Unitary, "Toxic-bert: Bert model fine-tuned for toxicity detection," https://huggingface.co/unitary/toxic-bert, 2024.

[72] Monologg, "Goemotions pytorch," 2024, accessed: [Date of Access]. [Online]. Available: https://github.com/monologg/GoEmotions-pytorch

[73] G. Research, "Goemotions," 2024, accessed: [Date of Access]. [Online]. Available: https://huggingface.co/datasets/google-research-datasets/go_emotions

[74] Y.-J. Kang and Y. Noh, "Development of hartigan's dip statistic with bimodality coefficient to assess multimodality of distributions," *Mathematical Problems in Engineering*, vol. 2019, no. 1, p. 4819475, 2019.

[75] Perspective API, "Perspective api," 2024, accessed: 2024-08-27. [Online]. Available: https://www.perspectiveapi.com/

[76] N. P. Madali, M. Alsaid, and S. Hawamdeh, "Social noise on social media and users perception of global warming," *Journal of Information & Knowledge Management*, vol. 22, no. 06, p. 2350050, 2023.

[77] S. Boulianne, M. Lalancette, and D. Ilkiw, ""school strike 4 climate": Social media and the international youth protest on climate change," *Media and Communication*, vol. 8, no. 2, pp. 208–218, 2020.

[78] C. Jia, M. S. Lam, M. C. Mai, J. T. Hancock, and M. S. Bernstein, "Embedding democratic values into social media ais via societal objective functions," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, pp. 1 – 36, 2023.

[79] W. E. S. Yu, "A framework for studying coordinated behaviour applied to the 2019 philippine midterm elections," in *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 2*. Springer, 2022, pp. 721–731.

[80] F. Souza and J. Filho, "Bert for sentiment analysis: Pre-trained and fine-tuned alternatives," 2022. [Online]. Available: https://arxiv.org/abs/2201.03382

[81] M. Wankhade and A. C. S. Rao, "Opinion analysis and aspect understanding during covid-19 pandemic using bert-bi-lstm ensemble method," *Scientific Reports*, vol. 12, no. 1, p. 17095, 2022.

[82] T. Bikku, J. Jarugula, L. Kongala, N. D. Tummala, and N. V. Donthiboina, "Exploring the effectiveness of bert for sentiment analysis on large-scale social media data," in *2023 3rd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2023, pp. 1–4.

[83] R. K. Das and D. T. Pedersen, "Semeval-2017 task 4: Sentiment analysis in twitter using bert," *arXiv preprint arXiv:2401.07944*, 2024.

[84] O. Renuka and N. Radhakrishnan, "Bert for twitter sentiment analysis: Achieving high accuracy and balanced performance," *Journal of Trends in Computer Science and Smart Technology*, vol. 6, no. 1, pp. 37–50, 2024.

[85] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the bert model," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 33, 2021.

[86] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "Covid-19 sensing: negative sentiment analysis on social media in china via bert model," *Ieee Access*, vol. 8, pp. 138 162–138 169, 2020.

[87] D. Fimoza, A. Amalia, and T. H. F. Harumy, "Sentiment analysis for movie review in bahasa indonesia using bert," in *2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*. IEEE, 2021, pp. 27–34.

[88] A. Sahoo, R. Chanda, N. Das, and B. Sadhukhan, "Comparative analysis of bert models for sentiment analysis on twitter data," in *2023 9th International Conference on Smart Computing and Communications (ICSCC)*. IEEE, 2023, pp. 658–663.

[89] T. Alsinet, J. Argelich, R. Béjar, and S. Martínez, "Measuring polarization in online debates," *Applied Sciences*, vol. 11, no. 24, p. 11879, 2021.

[90] M. Singh, S. Iyengar, and R. Kaur, "A multi-opinion based method for quantifying polarization on social networks," *arXiv preprint arXiv:2204.08697*, 2022.

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

[92] R. Interian and C. C. Ribeiro, "An empirical investigation of network polarization," *Applied Mathematics and Computation*, vol. 339, pp. 651–662, 2018.

[93] K. Huremović and A. I. Ozkes, "Polarization in networks: Identification–alienation framework," *Journal of Mathematical Economics*, vol. 102, p. 102732, 2022.

[94] R. Interian, R. G. Marzo, I. Mendoza, and C. C. Ribeiro, "Network polarization, filter bubbles, and echo chambers: An annotated review of measures and reduction methods," *International Transactions in Operational Research*, vol. 30, no. 6, pp. 3122–3158, 2023.

[95] Z. Lin, Z. Wang, Y. Tong, Y. Wang, Y. Guo, Y. Wang, and J. Shang, "Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation," *arXiv preprint arXiv:2310.17389*, 2023.

[96] A. Karimi, L. Rossi, and A. Prati, "Uniparma at semeval-2021 task 5: Toxic spans detection using characterbert and bag-of-words model," *arXiv preprint arXiv:2103.09645*, 2021.

[97] Helboukkouri, "character-bert," 2024, accessed: August 26, 2024. [Online]. Available: https://github.com/helboukkouri/character-bert

[98] M. Yoshida, K. Matsumoto, M. Yoshida, and K. Kita, "System to correct toxic expression with bert and to determine the effect of the attention value," in *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*. Springer, 2022, pp. 239–253.

[99] S. T. Luu and N. Nguyen, "Uit-ise-nlp at semeval-2021 task 5: Toxic spans detection with bilstm-crf and toxicbert comment classification," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, 2021. [Online]. Available: http://dx.doi.org/10.18653/v1/2021.semeval-1.113

[100] R. Singh, R. Kashyap, and V. Sharma, "Toxic comment analyzer using bert: A deep learning approach for toxicity detection," in *2023 Second International Conference on Informatics (ICI)*. IEEE, 2023, pp. 1–6.

[101] A. Wazzan and Y. Aldamen, "How university students evaluate the role of social media in political polarization: Perspectives of a sample of turkish undergraduate and graduate students," *Journalism and Media*, 2023.

[102] F. N. Selnes, "Fake news on social media: Understanding teens' (dis)engagement with news," *Media, Culture Society*, vol. 46, pp. 376 – 392, 2023.

# Acknowledgments

I would like to express my deepest gratitude to my supervisor, Prof. Alessandro Galeazzi, for giving me the opportunity to work on this project and for their invaluable guidance throughout this journey and a special thanks to Prof. Mauro Conti, whose support and insights were pivotal in starting this research process. Finally, I want to thank my family and friends, whose love and encouragement transformed the challenges of being an international student in a new country into an experience that was not only easier but also truly enjoyable.