

CS7IS4 - Group 6 - Midterm Paper

Unveiling the Spectrum of COVID-19 Perspectives from Social Media Based on Language Analysis

Rasika Vilas Burde
23331395
burder@tcd.ie

Xinyang Shao
23336365
shaoxi@tcd.ie

Rutuja Manohar Pate
23341465
pater@tcd.ie

Samarth Agrawal
23358805
agrawasa@tcd.ie

Pradosh Nataraj
23332074
pnataraj@tcd.ie

Qiang Han
23330515
hanq@tcd.ie

School of Computer Science and Statistics
Trinity College Dublin, The University of Dublin
February 2024

1 Introduction

Along with the rapid development of the internet and social media, the online community has become an important platform for people to share information, express opinions, and construct identities. As the foundation of social interaction, it is not only responsible for transferring information but also reflects the social attributes, cultural background, and psychological state.

During the COVID-19 epidemic, the government's vaccine roll-out policies and work-from-home measures caused extensive public discussion, and people expressed their stances and emotions on social media and online forums. In this process, the pronoun is the basic element in their language and the usage of pronouns may reveal the social stance, group identity, and emotional attitude of speakers. Therefore, to understand the division and unity of contemporary society and the relationship between the individual and collective, it is important to research the language usage of COVID-19 in the online community, especially the use of pronouns.

2 Research Question

The purpose of this study is to fill the gaps in the research on language usage during the epidemic period in online communities and explore how the language constructs social identities and expresses group positions by analyzing the use of pronouns in the discussion about the epidemic. Our study focuses on two scenes: The language usage difference among vaccination groups in online communities, which is based on the government's promotion of COVID-19 vaccines during the epidemic. Specifically, this study will focus on the following issues:

Are there significant differences in pronoun use in online communities between pro- and anti-vaccination groups during the COVID-19 pandemic? What social identities and group positions do these differences reflect?

By studying these issues, we want to have a deep understanding of how the usage of language reflects social divisions, personal identity and group belonging in online communities in the background of a global crisis. Also, the studies aim to explore how language interaction in online communities can be traded as an important method of understanding contemporary social activity and personal identity construction. We think the studies can support new perspectives and insights for understanding wider social and cultural phenomena through analysis of the language usage pattern under the COVID-19 epidemic background. At the same time, these studies provide important theoretical and empirical support for exploring the role of language in building and maintaining social relationships in online communities.

3 Literature Review

The novel approach of applying deep learning in community detection as stated in a 2023 study highlights the importance of using advanced algorithms to navigate the complexity of high-dimensional data within online communities[1]. This approach implies nuanced patterns of interaction and grouping within large datasets providing a methodology that establishes the foundation for investigating pronoun usage among different COVID-19-related forums on social media. In Complimentary the use of knowledge graphs for community detection in textual data[2] provides an advanced methodology for structuring and analyzing the information in any text and extracting meaningful insights from it. KGs organize data into interconnected nodes and facilitate a deeper understanding of the semantic relationships at a much deeper level within the online environment which is relevant to our paper's focus on pronoun deployment as an indicator of group formations during the COVID-19 pandemic.

Moreover, there have been studies conducted before that focus on certain sections of users and analyze patterns in their participation and behavior across similar communities on the same platform. they do conclude that people inherently tend to behave in a similar fashion holding the same point of view and opinions on a familiar topic across different groups. For reference, the study conducted by Kristen Engel[5] focuses on the participation of Reddit users in the QAnon conspiracy theory, characterizing the activities of early QAnon users on the platform, especially outside of QAnon-focused subreddits. The paper found that QAnon users participated in a diverse range of subreddits, often unrelated to QAnon, but most of their submissions were concentrated in subreddits sympathetic to the conspiracy theory while also being pro-Trump, in favor of hate speech, and anti-establishment. Although the studies lack depth in deriving patterns in casual conversations and acknowledge the lack of controlled comparisons and the potential gap between expert and novice discussions in the communities, the dynamics of language use and group identity might differ significantly in physical interactions as people may tend to pose a different image of themselves in real word to appeal to a wider audience or at times by posing a question to other people to try and gain their perspective on the said discussion before sharing their controversial views thereby affecting others.[6]

Additionally, the comparative study on community detection algorithms by Jenan, Moosa, Wasan, Shaker, Awad, and Tatiana, Kalganova [4] particularly within the context of COVID-19 data distribution provides valuable insight into the applications and efficiency of various analytical algorithmic approaches. The examination of various algorithms in tracing the narratives related to the COVID-19 virus transmission offers a promising comparative study with an emphasis on the role a language plays in forming and shaping public opinions and the group dynamics in response to the pandemic situations. This holistic analysis and comparative study lays a firm base for scrutinizing the broad spectrum of COVID-19-related perspectives on social media through language analysis.

On similar lines, the online discussions regarding the COVID-19 vaccine rollout present a complex mixture of communication strategies and linguistic choices that also reflect societal divisions in a broader context. Jialang Shi conducted a study [7], analyzed the tweets during the COVID-19 vaccine's initial rollout and presented the difference between pro- and anti-vaccine groups, where each group employed

different communication styles under the influence of group dynamics and social identities. Pro-vaccine discussions were characterized by positive messages with an approach of explanation and excitement, while those against vaccines expressed concerns and strategies of manipulation targeting pro-vaccine figures and policies. This division not only indicates the group's stand but also reflects how the platform policies provide moderation over the size and activities of the online communities and provide a fine-grained understanding of how social media platforms shape public health discussions

Further, the development and progress made in detecting non-referential pronouns presented by Shane, Bergsma, and David, Yarowsky [3] provide critical tools for the classification of referential and non-referential pronouns. This differentiation is vital for accurate interpretation of the context of pronouns in online communication which plays a critical role in ensuring precision of linguistic elements that signal group identity and also emotional attitude around the COVID-19 discussions.

Incorporating more insights from the research works of amira, Shaikh., Laurie, Beth, Feldman., Eliza, Barach., Yousri, Marzouki.[9] their study unveils a complex interplay of emotions, group dynamics, and social identity through lexical choices including pronouns that signal various psychological states and community behaviors.

Complementing the existing approaches, the Senti-COVID19 system [8] is a unique tool in the sentiment analysis space that was designed with the COVID-19 pandemic conversations on social media in focus. It used the Natural Language Toolkit and the VADER vocabulary to extract specific keywords and do reliable sentiment analysis, respectively. The Senti-COVID19 system made it easier to identify important impacts on public conversation by providing a thorough overview of sentiment patterns through its dynamic and multifaceted visual interface. Case studies demonstrate this system's proficiency in processing large-scale sentiment analysis and highlight how well it reflects the societal impact of COVID-19 and the subtle changes in public mood across social media platforms.

Gaining insights from these methodologies our paper will deploy sophisticated pronoun identification techniques and improve upon the current understanding of how language reflects as well as constructs social identities and groups amidst a global crisis. This literature review not only emphasizes the interdisciplinary approaches employed in analyzing online communities but also paves the way for investigating language nuances that classify and characterize public discussions on vaccination and COVID-19-related policy implementation that takes place on social media platforms.

4 Research Methodology

In our research technique, we want to look into how online communities, particularly those focusing on COVID-19, arise and are defined by talks about a variety of topics. Our investigation extends to comprehending language nuances, such as the use of pronouns "I," "we," "our," and "they," as well as other suggestive phrases that indicate group connections and separations inside the r/DebateVaccines and r/Coronavirus subreddits. The research is intended to determine whether digital forums that serve various functional demands, such as mutual support versus enjoyment, influence how people and groups develop and engage within these communities.

I Data Collection

Firstly, during the data collection, we are going to use the **Requests** library to collect the data from the COVID-19-related subreddits. We extended our scraping efforts to include specific subreddits, notably r/DebateVaccines and r/Coronavirus which was done using the Requests library. This script is carefully designed for browsing the Reddit application, which is also combined with corporate paging functionality to access and collect data from multiple pages systematically. We add the user agent header to our request

data, which is a key element in ensuring access to the Reddit API guidelines and maintaining respectful access behaviour, to simulate the actions of a normal user. To simulate matching user requests to the server, the script adds timestamps between requests, which we can use the `time()` method in the API to potentially lighten the load on the Reddit server. The script is configured to crawl around 5000 posts at appropriate intervals and can be configured to set parameters according to the user's needs, such as the data crawled per page, and information about the data viewed to collect more data, thus ensuring a comprehensive dataset for analysis.

II Data Cleaning

We are going to use the pandas library to analyze the Reddit data during the data cleaning process. This is the first step and is to confirm that the data we collected is to identify and remove duplicate values from affecting the results. Handling the missing data is another important key, which is based on available information or excluding incomplete records together. Text normalization involves converting all the text to a consistent case, basically, the lower case, removing punctuation, and standardizing expressions to reduce variability in the dataset. This step is essential because it provides an important foundation for the text analysis we are going on.

III Tokenization

Tokenization is the process of dividing the text into a sequence of tokens, which might be words, phrases, symbols, or other meaningful elements. This step is the foundation for NLP processing, because it transforms the unstructured algorithms that can be interpreted by NLP, and can easily understand the text's linguistic structure. In our implementation, we use the natural language library(NLTK) for tokenization, which is a pivotal step for preparing text from linguistic analysis. NLTK is famous for its widened natural processing and provides `word_tokenize` and `sent_tokenize` functions. This tokenization process is not only fundamental for parsing and semantic analysis but also crucial for accurately identifying and analyzing linguistic patterns within the textual data from Reddit discussions. Our choice of NLTK over other libraries was influenced by its extensive documentation, ease of use, and its wide range of features tailored for in-depth linguistic research. more helpful to convert the complex context data to understandable phrases and words.

IV Frequency and Contextual Analysis

We use scikit-learn for TF-IDF analysis to identify key occurrences or words by analyzing the relative importance of key belongings or words in different contexts. TF-IDF (word frequency-inverse document frequency) is a statistical measure used to assess the importance of a word in a material relative to a corpus. **scikit-learn** provides a straightforward implementation for tf-idf analysis. In the TF-IDF analysis, gensim and NLTK provide a straightforward implementation to make sure that these essential terms can be further analyzed. For the collaboration, both gensim and NLTK are valuable. Gensim are professional from a large language context, allowing word2Vec to capture word associations and similarities. NLTK helps identify collocations, i.e., pairs of words that occur more frequently than by chance, thus enriching the understanding of the textual context. All of these methods can help us to more understand the associations of the worlds in the context. The converter library provides the access to the BERT Model and can further analyse the context, which are based on around each world. So, these tools can actually improve our use in context libraries to analyse the word frequency and context understanding, providing the insights into language views.

Our main study revolves around analysis of pronoun use -"I," "we," "our," and "they"- into our research which significantly enhances our understanding of identity dynamics in online groups. We are able to thoroughly assess the frequency and context of these pronouns by utilizing the complex text processing and part-of-speech tagging capabilities of NLTK and spaCy. By using this analytic approach, we can

identify patterns of group dynamics and self-representation, which gives us insights into how people and groups interact, align, and set themselves apart in digital spaces. By providing a comprehensive perspective of the social structures at work in online interactions, this approach is helpful in determining involvement levels, solidarity, and divisions.

V Sentiment Analysis

To evaluate emotional content in subreddit talks, sentiment analysis is performed using the NLTK and TextBlob libraries. VADER in NLTK calculates sentiment scores, which indicate whether a text is positive, negative, or neutral. TextBlob is also helpful in determining subjectivity and polarity. These scores are critical for assessing community reactions and dynamics, providing insights into the dominant sentiments across various discussion threads, and greatly contribute to our understanding of online community behavior and engagement patterns in the context of COVID-19 discourse.

VI Comparative Analysis

We used comparative analysis to investigate differences in linguistic styles, attitudes, and thematic concerns across several Reddit communities, such as r/DebateVaccines, that discuss COVID-19. Using TF-IDF and LDA algorithms from the gensim package, we identify distinct lexical items and subjects common to each subreddit. TF-IDF identifies phrases unique to each group, emphasizing thematic variety. Concurrently, LDA makes it easier to extract dominating subjects, giving you a better understanding of the content's focus. This dual method deepens our understanding of how COVID-19 discourse varies, highlighting each community's distinct perspective and shared issues.

VII Visualization

In order to make patterns and clusters visually apparent, we use t-SNE for dimensionality reduction in order to project high-dimensional data into a two-dimensional space. Then, using the linguistic characteristics and themes found in our research, K-means clustering is utilized to put related data points in groups. These techniques make it easier to identify unique groups and conversation topics within the subreddits and offer a visual representation of the intricate data structure that textual research alone might not be able to provide.

VIII LIWC

After completing initial pronoun and sentiment analysis with Python tools, we use LIWC to delve further into the psychological and emotional underpinnings of subreddit communications. LIWC organizes words into psychologically relevant groupings, improving our understanding of community dynamics by revealing information about emotional tone, cognitive style, and social orientation. This layered method, which combines classic NLP analyses with LIWC, enables a more sophisticated interpretation of how people express and position themselves in online debates, enriching our understanding of the complex social fabric of online communities.

5 Discussion

At the beginning of our research project, group members had a preliminary discussion about the topic of “how people establish different groups by using pronouns in the context of the COVID-19 epidemic”. This discussion aims to clarify our research direction, determine the research question, and plan the basic framework of the research.

First, we discussed the motivation and importance of choosing this topic. The online community is an important platform for people to exchange information and express opinions, the group members

generally agreed that language usage pattern reflects the changes in social attitude and group identity in the effect of the global epidemic. Especially the use of pronouns, which can reveal the identity stance and attitude of speakers. Therefore, we believe this is an important topic which is worthy of deep study.

Next, we discussed the research methodology and strategy of data collection. We decided to use internet research and content analysis at the start stage of research. Group members gave different advice on collecting related posts from social media and online forums, like which platform should we choose, how to identify and detect the usage of pronouns, and how to ensure the representativeness and effectiveness of data. We decided to assign the task based on the interest of each group member to ensure the efficient progress of research work.

Additionally, we also discussed the potential challenges and the solution strategy. Some team members worried that the large scale of data may cause too much analysis work, and some team members focused on the accuracy and fairness of the research result. In order to meet these challenges, our group established clear data filtering criteria chose automated tools to aid analysis, and had regular group discussions to monitor the research progress.

Finally, we have confirmed the next plan of research, which includes assigned tasks to collect and analyze preliminary data, continue the regular meetings to follow up on research progress, and solve problems. Through the existing discussions, our group members had a clear understanding of the research topic and were full of expectations for the upcoming research work

References

1. (2023). A Novel Community Detection Algorithm Based on Deep Learning Algorithm. Communications in computer and information science,doi: 10.1007/978-981-99-0301-6_30.
2. (2022). Knowledge Graphs for Community Detection in Textual Data. doi: 10.1007/978-3-031-21422-6_15
3. Shane, Bergsma., David, Yarowsky. (2011). NADA: a robust system for non-referential pronoun detection. doi: 10.1007/978-3-642-25917-3_2
4. Jenan, Moosa., Wasan, Shaker, Awad., Tatiana, Kalganova. (2021). Intelligent Community Detection: Comparative Study (COVID19 Dataset). doi: 10.1007/978-3-030-77246-8_19
5. Kristen Engel, Yiqing Hua, Taixiang Zeng, and Mor Naaman. 2022. Characterizing Reddit Participation of Users Who Engage in the QAnon Conspiracy Theories. Proc. ACM Hum.-Comput. Interact. 6, CSCW1, Article 53 (April 2022), 22 pages. <https://doi.org/10.1145/3512900>
6. Dow, B. J., Johnson, A. L., Wang, C. S., Whitson, J., and Menon, T. 2021. The COVID-19 pandemic and the search for structure: Social media and conspiracy theories. Social and Personality Psychology Compass, 15(9). <https://doi.org/10.1111/spc3.12636>
7. Blane JT, Bellutta D, Carley KM. Social-Cyber Maneuvers During the COVID-19 Vaccine Initial Rollout: Content Analysis of Tweets. J Med Internet Res. 2022 Mar 7;24(3):e34040. doi: 10.2196/34040. PMID: 35044302; PMCID: PMC8903203.
8. Kaushal A, Mandal A, Khanna D, Acharjee A. Analysis of the opinions of individuals on the COVID-19 vaccination on social media. DIGITAL HEALTH. 2023;9. doi:10.1177/20552076231186246
9. Samira, Shaikh., Laurie, Beth, Feldman., Eliza, Barach., Yousri, Marzouki. (2017). Tweet Sentiment Analysis with Pronoun Choice Reveals Online Community Dynamics in Response to Crisis Events. 345-356. doi: 10.1007/978-3-319-41636-6_28

Statements of Contribution

Student ID	Full Name	Role	Nature Of Contribution
23331395	Rasika Vilas Burde	Monitor	As the monitor for Group 6, I was in charge of ensuring that each team member, including myself, submitted an original summary of a relevant scholarly paper, providing the framework for our complete synthetic review. I actively participated in weekly meetings and engaged in discussions based on the most recent study findings. In addition, I helped to find appropriate datasets for our analysis and was instrumental in designing the methodology section, which required a good amount of research and the use of advanced analytical techniques inclined to our study's emphasis.
23341465	Rutuja Manohar Pate	Recorder	As the recorder of group 6, I was responsible for recording the minutes of every meeting and checking if the tasks decided in the previous meetings and the current meeting were being completed and synchronized. Additionally, I ensured my regular participation in the group meetings and contributed to the group discussions by demonstrating ideas acquired by researching relevant and current trends. I made a significant contribution to the literature review by providing insights on state-of-the-art methods for the classification of referential and non-referential pronouns, community detection using deep learning algorithms, and the use of knowledge graphs in community detection, as well as a comparative analysis of community detection algorithms. All my group members have also shown great enthusiasm and have equally contributed to the midterm paper.
23358805	Samarth Agrawal	Ambassador	Being the ambassador of group 6, I met with other group members namely group 1 and group 7, and exchanged ideas with them. I did extensive research on Covid communities and papers related to them. As for the midterm paper, everyone contributed equally to the paper.

23336365	Xinyang Shao	Chair	In our group project, I took the role of chair and was responsible for arranging meeting times that all members could attend, formulating the meeting agenda (soliciting opinions from other members), chairing the meeting, ensuring that the monitor understands the articles that each member (including myself) reads and summarises each week, and is responsible for the group's communication with Communication between lecturers. In addition to my responsibilities, I searched for literature, data sets, etc related to the research direction. I also actively participate in discussions at every meeting and provide valuable perspectives. In the midterm paper, everyone contributes equally to the paper in research questions, literature review, methodology and discussion.
23332074	Pradosh Nataraj	Verifier	As the verifier for Group 6, my role was to ensure that every member of the group fulfilled their responsibilities effectively. From the outset, I observed that the group was highly engaged and active. My contribution included guiding the group through detailed discussions about our topic, assisting in locating the appropriate datasets, and finding relevant research papers.m For our mid-term paper, we adopted a collaborative approach where each member was assigned specific tasks, followed by a review process by others in the group. My task was to conduct a literature review, which involved researching and analysing various papers pertinent to our topic.
23330515	Qiang Han	Accountant	In our group project, I took the role of accountant responsible for all the group meeting details, including the group date, the duration, the specific content of the meeting, and the contributions of each member. During the every week meeting, I actively participated and engaged in discussion project topics. In addition, I did a lot of research for our project methodology and provided prospective ideas. As for the mid term paper, everyone contributed equally to the paper.

Member 1: Rasika Vilas Burde



Member 2: Samarth Agrawal



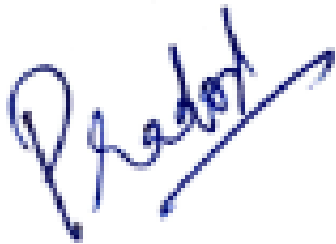
Member 3: Rutuja Manohar Pate



Member 4: Xinyang Shao



Member 5: Pradosh Nataraj

A stylized, handwritten signature in blue ink. The signature appears to be 'Pradosh' followed by a large, sweeping flourish that extends to the right.

Member 6: Qiang Han

A handwritten signature in black ink. The signature is written in a cursive style, with the first name 'Qiang' and the last name 'Han' clearly visible.