Aspect-Based Sentiment Analysis On LGBTQ+ Tweets

Abstract

It is commonplace in today's world for people to discuss their feelings and views on a wide range of issues on social media. X, formerly Twitter is a major platform for opinion sharing in most parts of the world. Discussions about LGBTQ+ has been on the rise recently, especially on X which motivated this research. This study classifies 32,456 tweets with #lgbt to see public views on the LGBTQ community using Sentiment Analysis approaches. TextBlob is employed to obtain the polarity and subjectivity of tweets which reveal that 49.8% of the extracted tweets are positive, 25.7% are neutral and 24.5% are negative. Word Tokenize from NLTK is used to extract the most occurring words for Aspect Based Sentiment Analysis (ABSA). The most occurring words across all the tweets are "lgbt", "people", "community", "like", "anti", "pride", "right", "gay", "support", "say" and "significant". Key findings highlight that LGBTQ community involvement is stronger on social media with almost half of the extracted tweets giving positive sentiments. Major aspects in discussion are support for the people, community rights, and support gay.

Keywords
Aspect-Based Sentiment Analysis
Social media
LGBTQ+
Natural Language Processing
Aspect Extraction

Introduction

In today's world, social media serves as an integral vehicle[6] and as an online community[7] for seeking and sharing information, news, views, opinions, perspectives, ideas, awareness, comments, and experiences on various topics, such as pandemics, global affairs, current technologies, recent events, politics, family, relationships, and career opportunities, just to name a few[8]. Twitter has been identified as a powerful tool for disseminating real time information to promote public awareness. Approximately, 2.95% of the world's population uses Twitter[18]. Examining evolving discussion regarding the outbreak are useful to determine the components that are informing public views. During COVID-19, Twitter data was utilized for scientific research to observe users' rising concerns, spread of misinformation, and overall sentiment[9]. Nonetheless, the qualities of severe ineffectiveness, irrationality, and uniformity are ascribed to public opinion extracted from social media. Despite these drawbacks it is understood that social media is essential to the social and technological framework that allows people to remain connected during emergencies as it has emerged as major platform for communication[10].

The goal of sentiment analysis, commonly referred to as opinion mining, is to automatically identify the opinions expressed in text[2]. Sentiment Analysis has historically been interpreted as the degree to which someone has expressed a favorable, unfavorable, or neutral view of an event[3]. Since the beginning of the previous ten years, researchers have worked to collect, quantify, and measure the dynamic public attitudes using a variety of tools, methods, and strategies, making SA one of the research topics that is expanding quickly[4] [5]. Aspect-based sentiment analysis (ABSA) is a fine-grained sentiment analysis task; its purpose is to identify the different aspects of a text and judge their corresponding sentiment polarity[1]. Aspect-based sentiment analysis (ABSA) is a natural language processing (NLP) technique used to analyze sentiment expressed towards specific aspects or features of a given topic.

Sexual orientation refers to those you are attracted to and desire to be involved with[11]. Several people nowadays are classified as having different sexual orientations. The terms lesbian, gay, bisexual, transgender, and queer (LGBTQ) refer to a person's sexual orientation or gender identity[12]. There are more than 11 million lesbian, gay, bisexual, and transgender (LGBT) adults in the United States[13]. Social media has become a mainstream channel of communication and has grown in popularity. Social media facilitates people's belonging to and exchanging information within LGBT communities by allowing users to transcend geographic barriers in online spaces with the limited risk of being "outed" [14]. Compared to heterosexual respondents, LGBT users are more likely to have accounts on social media websites, access social media daily, and make frequent use of the internet[15].

The use of social media platforms has become an essential tool for the LGBTQ community, as well as its allies, advocates, and even those who oppose it. It provides a space for individuals to express their opinions, share personal experiences, and engage in conversations about topics like identity, representation, rights, and equality. According to a survey, 80% of LGBT Americans use social networking websites, and about four in ten LGBT adults have revealed their sexual orientation or gender identity on social networking sites[16]. While these platforms allow for a wide range of voices to be heard, they also expose users to various attitudes and emotions. These can range from acceptance and solidarity to discrimination and hostility. Sentiment analysis combines natural language processing with data science to offer a systematic way to study the emotional aspects of written content. By applying sentiment analysis techniques to a collection of LGBTQ-themed tweets from X , we can gain insights into the prevailing attitudes and emotional tones that characterize these discussions. This information can help us understand how society perceives issues related to LGBTQ individuals while also highlighting any emerging trends within this discourse.

This research paper aims to contribute to the existing literature on sentiment analysis and LGBTQ studies by conducting an aspect-based sentiment analysis of tweets related to LGBTQ topics from X. By employing various techniques, this study will help us uncover sentiment variations of distinct aspects within the LGBTQ discourse. Subsequent sections present the methodology used for data collection, preprocessing, aspect extraction, and sentiment classification. The results of the sentiment analysis are presented, followed by a discussion of the implications of these findings for both LGBTQ advocacy and sentiment analysis methodology. Furthermore, I acknowledge the limitations of the study and suggest avenues for future research to continue exploring societal sentiments surrounding LGBTQ topics.

Methodology

Data Collection and Preparation

Dataset

To conduct an aspect-based sentiment analysis on LGBTQ-related tweets, a publicly available dataset was sourced from Kaggle[17]. The dataset has 9 columns which includes time of tweet, date, number of retweets and likes as well as the language. The dataset contains 32, 456 datapoints. It is worthy to state that the tweets were scraped in August 2022. The focus was on tweets that contained LGBTQ-related keywords, hashtags, and mentions.

Data Cleaning

The raw tweets often contain shorthand text, abbreviations, and irregularities commonly used on social media platforms. To enhance the quality of the text for analysis, Beautiful Soup was used. This allowed the conversion of shorthand and informal language into their complete forms, to improve the accuracy of subsequent analyses.

The dataset contained noise in the form of special characters, URLs, emojis, and other non-essential elements that could impact sentiment analysis. Regular Expression patterns were applied to clean the text by removing these unwanted elements. This process cleaned the text, ensuring that only meaningful content was used for analysis.

To ensure consistent word representation, SpaCy, a natural language processing library, was employed for lemmatization. Lemmatization reduced words to their base forms, accounting for variations in tense, plurality, and conjugation. This step enhanced the accuracy of subsequent text analysis, particularly in sentiment classification.

Stopwords, commonly used words like "the," "and", "is," etc., were removed from the text using word tokenization and predefined lists of stopwords. Word tokenization also split the text into individual words or tokens. This process reduced the dimensionality of the data while retaining its meaningful content for sentiment analysis. Figure 1 shows the WordCloud generated after the cleaning the data

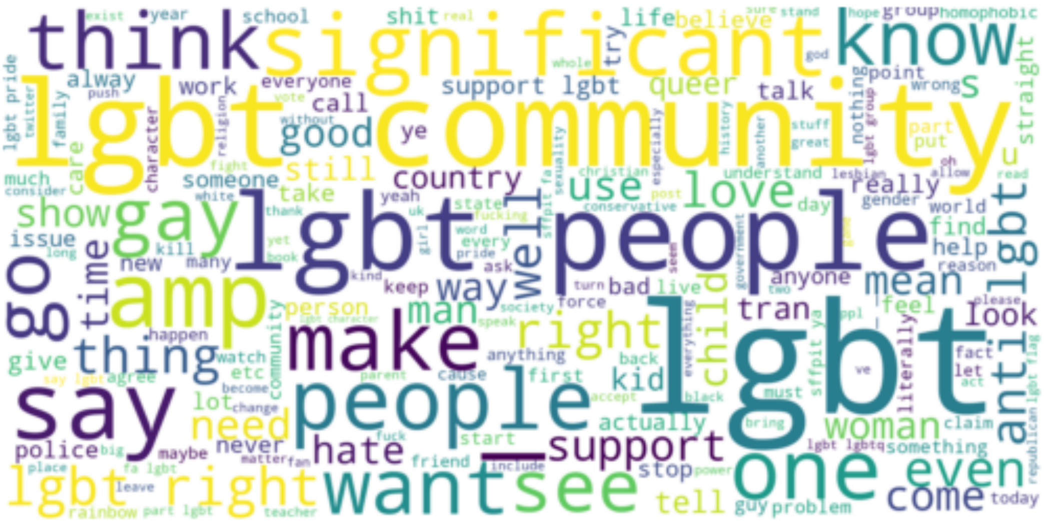


Figure 1 : WordCloud of entire Dataset

Sentiment Classification

The decision to use TextBlob's sentiment analysis module was driven by its accessibility and user-friendly nature. With its straightforward interface, TextBlob empowered me to conduct sentiment analysis efficiently, even without extensive expertise in natural language processing. This choice aligned with the research's aim to delve into sentiment patterns while keeping the process manageable and insightful.

I leveraged TextBlob's polarity scores to capture the essence of each tweet's sentiment on a numerical scale. By obtaining polarity scores ranging from -1 to 1(-1 represents extremely negative sentiment, 0 represents neutrality and 1 represent extreme positivity), I could gauge the sentiment's polarity, whether it leaned towards a negative, positive, or neutral expression. These scores provided me with a quick understanding of the emotional context of each tweet, allowing for easy comparisons across aspects and categories. Figure 2 displays the polarity and subjectivity of the tweets.

Incorporating subjectivity scores into the analysis further enriched my insights. These scores helped differentiate between tweets that expressed personal opinions and those that conveyed factual information. This dimension added depth to my understanding of how emotionally invested individuals were in their tweets, thereby uncovering the intensity of their sentiments.

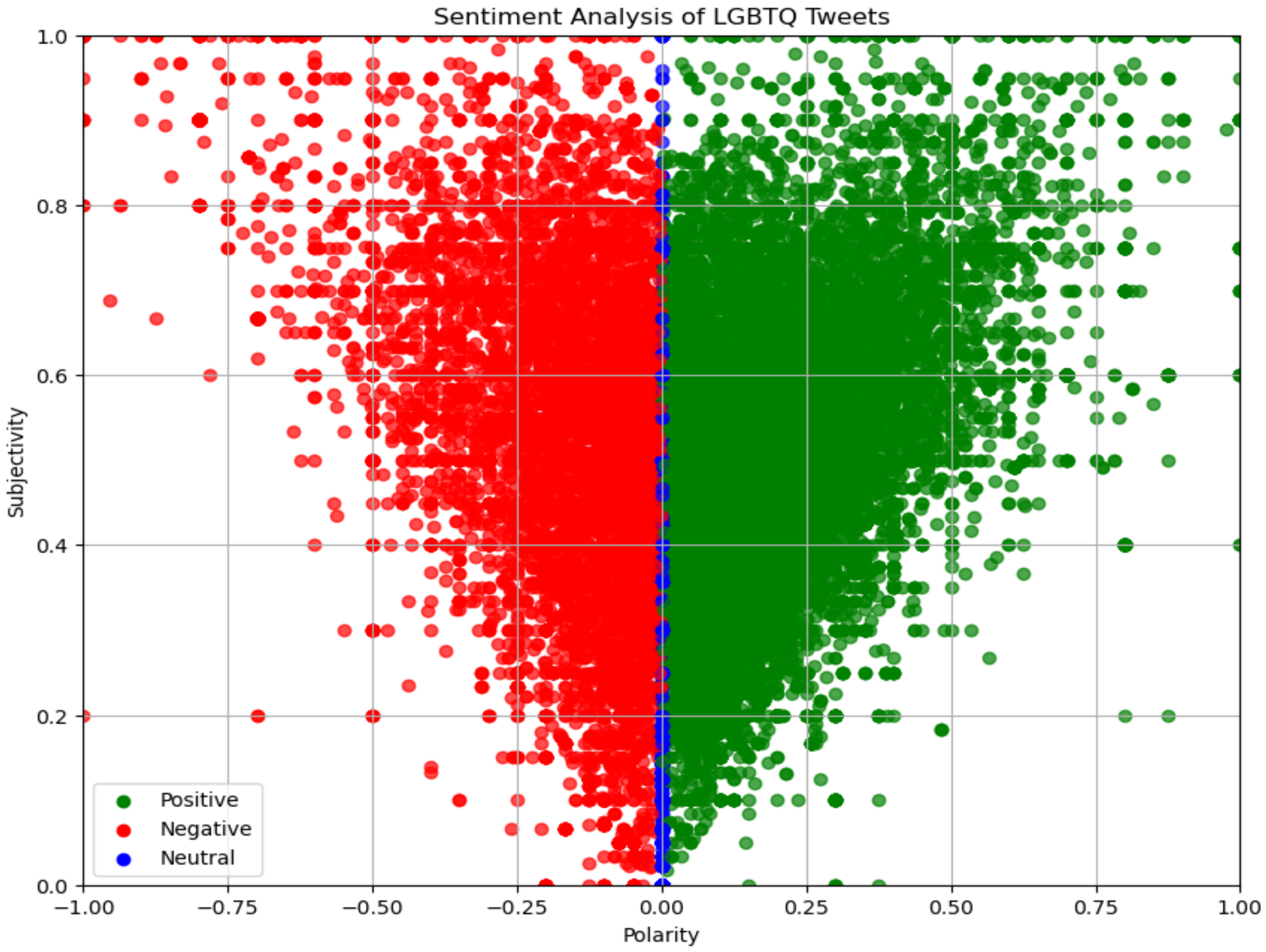


Figure 2 : Sentiment Classification

Aspect Identification

The choice to employ word tokenization for aspect identification was rooted in its adaptability and precision. Word tokenization served as a versatile technique that allowed for the extraction of keywords and phrases relevant to each aspect. This approach resonated with the research's objectives to tailor the aspect identification process to the unique nuances of LGBTQ discussions. When considering alternative approaches for aspect identification, Named Entity Recognition (NER) emerges as a plausible option to contrast with word tokenization. NER is a technique used to identify and categorize named entities, such as names of people, organizations, locations, and other specific terms within text.
Word tokenization offers the advantage of tailoring keywords to specific aspects, facilitating personalized analysis. NER automates aspect identification but might lack the nuance required for precise analysis. Word tokenization allows for precise identification of terms directly related to an aspect. NER, on the other hand, can identify contextually relevant named entities but might not capture terms specific to the aspect.

Results

Sentiment Analysis

The sentiment analysis unveiled a diverse distribution of sentiments expressed within the dataset. Out of the analyzed LGBTQ tweets, approximately 49.8% exhibited positive sentiments, indicating a significant prevalence of optimism, support, and enthusiasm within the discourse. This positivity underscores the online platform's role as a space for fostering connections and promoting affirmative narratives.

On the other hand, 24.5% of the tweets conveyed negative sentiments. This subset encompassed expressions of criticism, frustration, and concerns that reflect the challenges and obstacles faced by the LGBTQ community. These negative sentiments highlight the ongoing struggles and the need for continued advocacy and awareness. Figure 3 shows the distribution of sentiments expressed in a pie chart.

25.7% of the tweets were categorized as neutral, signifying the presence of factual statements, news updates, or content that does not convey a strong emotional stance. These neutral tweets contribute to a balanced discourse by providing information and context without a pronounced emotional charge.

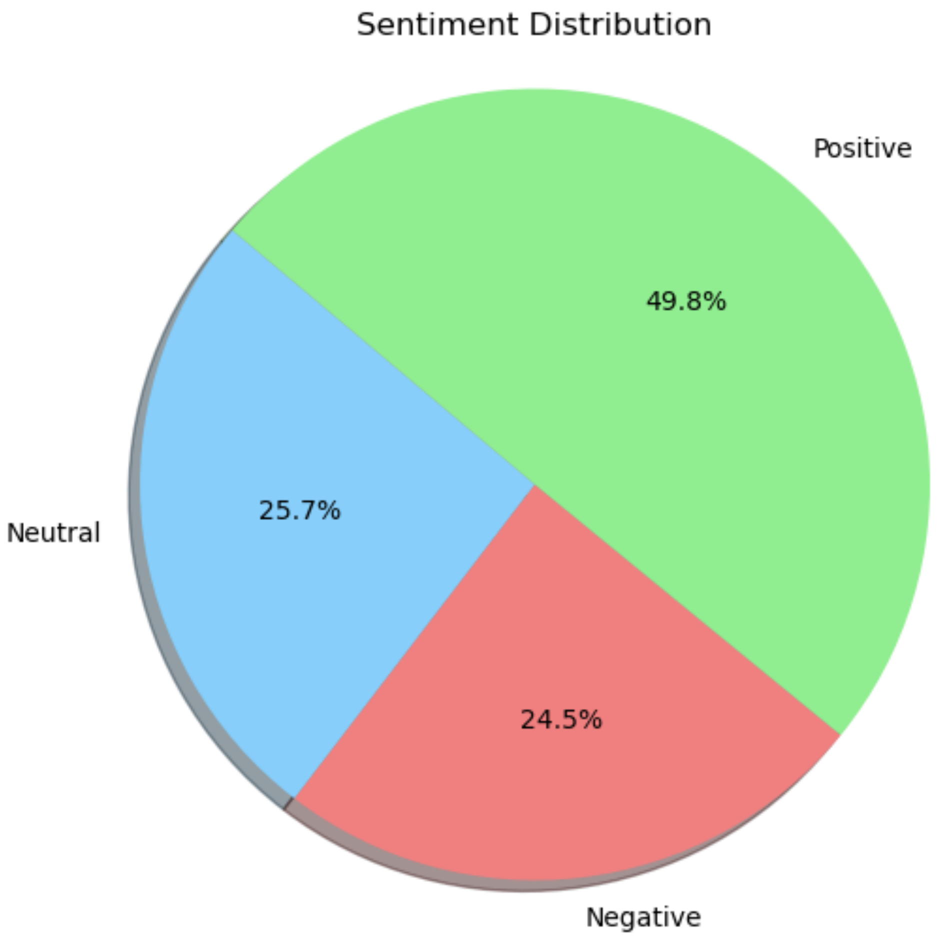


Figure 3 : Sentiment Distribution

Aspect Analysis

The analysis of the most frequently occurring words in the dataset revealed significant insights into the prevalent themes and language patterns. Among these, the term "lgbt" dominated with 30,501 occurrences, reflecting the central focus of the conversation. The occurrence of words like "people," "community," and "gay" underscores the emphasis on individuals' experiences, the sense of belonging, and diverse identities within the LGBTQ community.

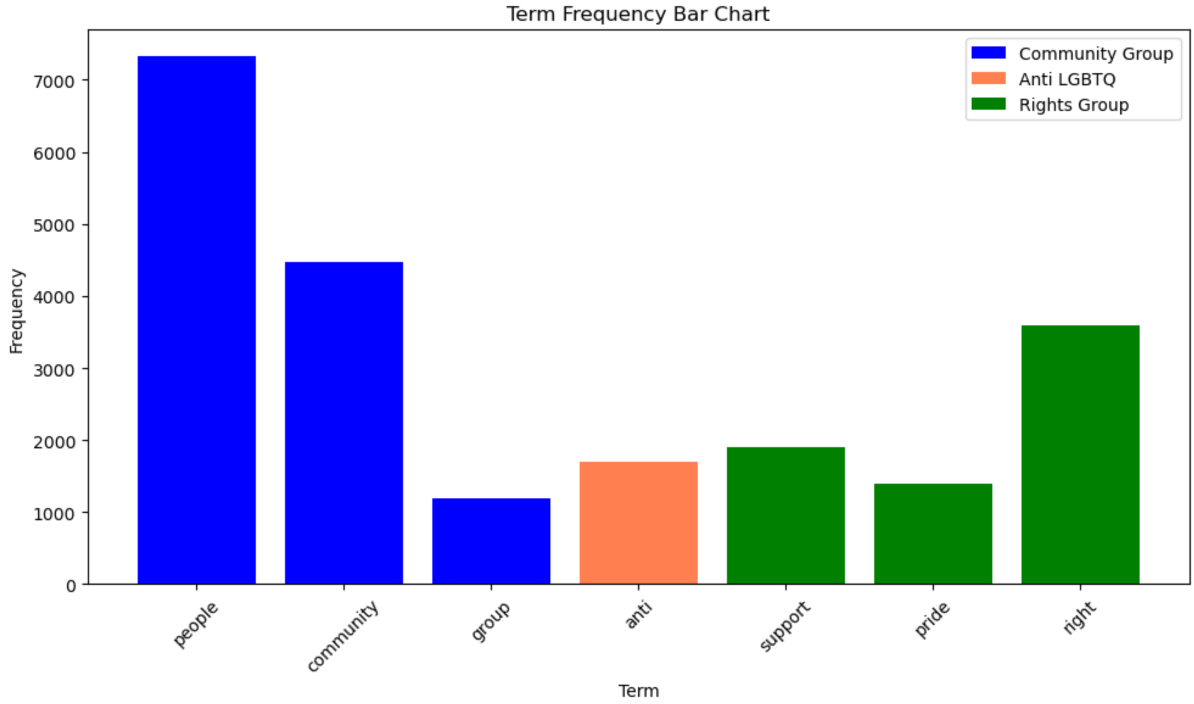


Figure 4: Most occurring words by Group

Positive Aspects

Certain aspects exhibited a predominantly positive sentiment. For instance, discussions around LGBTQ pride events and celebrations resonated with sentiments of joy, support, and empowerment. The term "pride," occurring 1,389 times, emphasizes the positive sentiments associated with events that celebrate the LGBTQ community's achievements and visibility.

Negative Aspects

Themes related to challenges faced by the LGBTQ community, such as discrimination and anti-LGBTQ sentiment, often garnered negative sentiments. The occurrence of terms like "anti" (1,697 occurrences) and "hate" (1,395 occurrences) highlights the prevalence of negative emotions associated with instances of prejudice and bias.

Neutral Aspects

Neutral sentiments were often observed in discussions that provided factual information, news updates, or general observations. The prominence of terms like "like" (4,169 occurrences) and "say" (3,083 occurrences) suggests the presence of factual statements and discussions that contribute to a balanced discourse.

Figure 6 and Figure 5 provide a snapshots of the themes and language prevalent in the LGBTQ discourse on Twitter. The repeated occurrence of words related to rights, community, support, and identity reflects the multifaceted nature of discussions within the LGBTQ community. These terms represent both the positive strides and challenges encountered by the community, offering a glimpse into the diverse sentiments expressed in digital conversations.

Figure 4 shows a bar chart that depicts the frequency of specific terms within the LGBTQ discourse on Twitter. The terms "people," "community," and "group" are clustered together and highlighted in blue. This grouping suggests a recurring emphasis on the LGBTQ community's collective engagement and collaborative efforts. The frequency of these terms implies an active involvement of individuals in group activities, underscoring the sense of belonging and unity within the community.

The terms "support," "pride," and "right" are visualized in green, implying a strong presence of discussions related to advocacy for LGBTQ rights. The considerable frequency of these terms reflects a prominent discourse surrounding support for LGBTQ individuals and their rights. The term "pride" highlights the celebration and affirmation of LGBTQ identities.

The term "anti," marked in coral, stands out distinctly from the rest. Its frequency indicates the prevalence of discussions around anti-LGBTQ sentiments and actions.

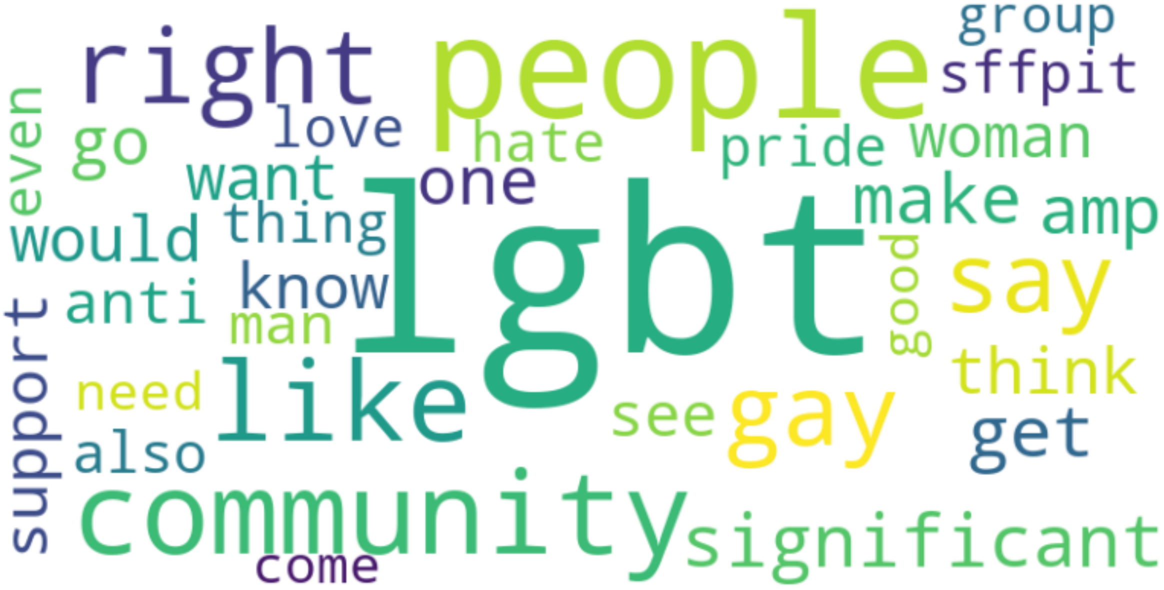
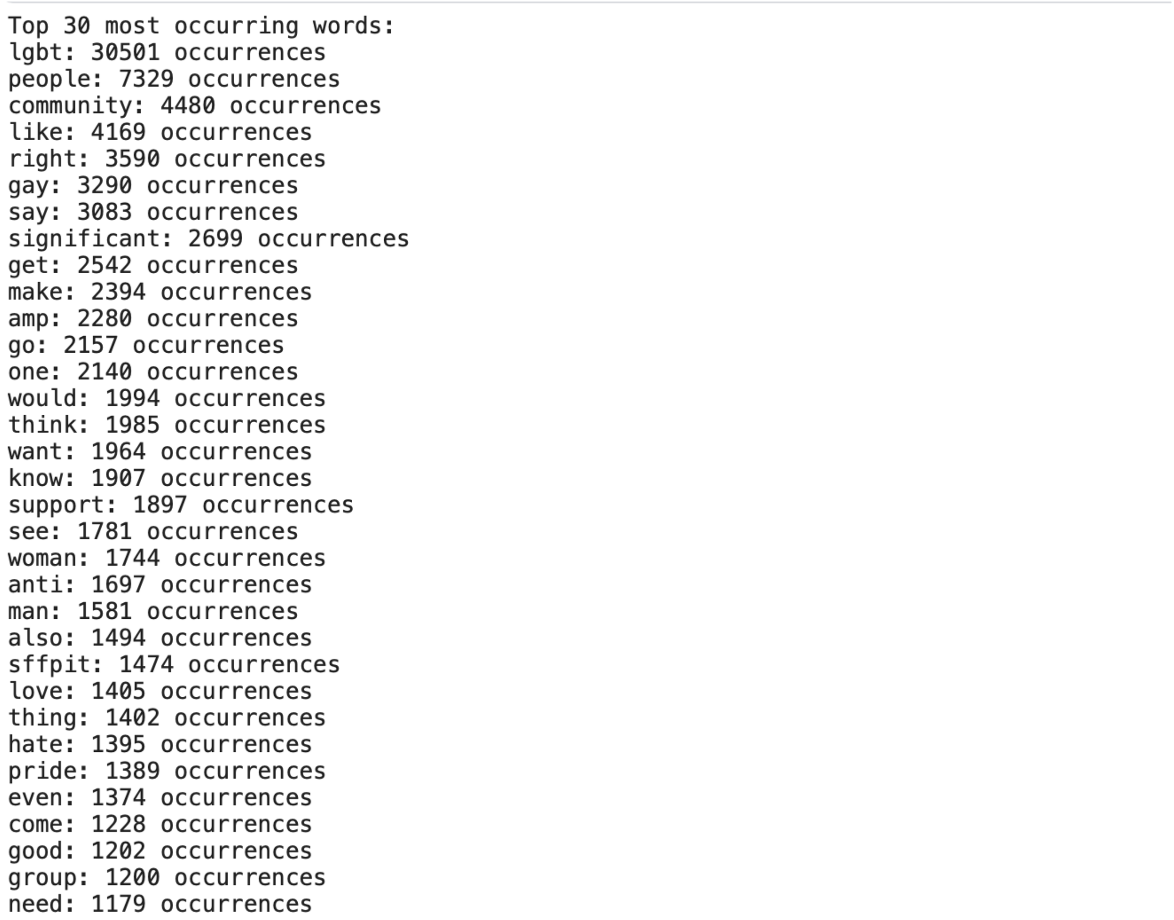


Figure 5 : WordCloud of 30 most occurring words



```
Top 30 most occurring words:
lgbt: 30501 occurrences
people: 7329 occurrences
community: 4480 occurrences
like: 4169 occurrences
right: 3590 occurrences
gay: 3290 occurrences
say: 3083 occurrences
significant: 2699 occurrences
get: 2542 occurrences
make: 2394 occurrences
amp: 2280 occurrences
go: 2157 occurrences
one: 2140 occurrences
would: 1994 occurrences
think: 1985 occurrences
want: 1964 occurrences
know: 1907 occurrences
support: 1897 occurrences
see: 1781 occurrences
woman: 1744 occurrences
anti: 1697 occurrences
man: 1581 occurrences
also: 1494 occurrences
sffpit: 1474 occurrences
love: 1405 occurrences
thing: 1402 occurrences
hate: 1395 occurrences
pride: 1389 occurrences
even: 1374 occurrences
come: 1228 occurrences
good: 1202 occurrences
group: 1200 occurrences
need: 1179 occurrences
```

Figure 6 : Top 30 most occurring words

Discussions

Comparison with Existing Literature

Appiahene et al. (Appiahene, Varadarajan, Zhang, & Afrifa, 2023) presented a study which aims to use natural language processing (NLP) and machine learning (ML) approaches to assess the experiences of LGBTQ+ persons. To train the data, the study used lexicon-based sentiment analysis (SA) and six distinct machine classifiers, including logistic regression (LR), support vector machine (SVM), naïve bayes (NB), decision tree (DT), random forest (RF), and gradient boosting (GB). Individuals are positive about LGBTQ concerns, according to the SA results; yet prejudice and harsh statements against the LGBTQ people persist in many regions where they live, according to the negative sentiment ratings. Furthermore, using LR, SVM, NB, DT, RF, and GB, the ML classifiers attained considerable accuracy values of 97%, 96%, 88%, 100%, 92%, and 91%, respectively. The performance assessment metrics used obtained significant recall and precision values.

Additionally, Dsouza et al. (Dsouza, et al., 2023) introduced a model of stigma communication, which maps and determines the mpox stigma on Twitter among LGBTQ+ (Lesbian, gay, bisexual, transgender, queer and more) community. The tweets that contained the terms '#monkeypox', '#MPVS', '#stigma', and '#LGBTQ+' and were published between May 01, 2022, and Sept 07, 2022 were extracted. For sentiment analysis, the VADER, Text Blob, and Flair analysers were implemented. This study evaluated the dynamics of stigma communication based on the "model of stigma communication". A total of 70,832 tweets were extracted, from which 66,387 tweets were passed to the sentiment analyser and 3100 tweets were randomly selected for manual coding. Consolidated Criteria for Reporting Qualitative Research (COREQ) criteria was adopted to report this study.

Implications and Applications

The insights derived from this study hold implications for both academia and society at large. Academically, the study contributes to a deeper understanding of sentiment dynamics within the LGBTQ discourse on Twitter, showcasing the multifaceted nature of opinions and emotions. This nuanced view can inform research in areas such as social psychology, communication studies, and digital sociology. On a broader scale, the study's findings provide insights for LGBTQ advocacy groups, policymakers, and social media platforms aiming to create more inclusive environments and address challenges faced by the community.

Limitations

It is important to acknowledge the limitations of this study. The sentiment analysis might not capture the full complexity of emotions expressed in text, as sarcasm, irony, and context can affect interpretation. Additionally, the dataset's representativeness and potential biases could influence the accuracy of sentiment analysis. The aspect identification process also relies on predefined keywords, potentially overlooking emerging themes. Lastly, the sentiment analysis might not account for nuanced cultural differences or evolving language trends.

Future Work and Improvements

Future research could overcome limitations by employing advanced sentiment analysis techniques that consider context and tone more comprehensively. Natural language processing models could be fine-tuned to the specific nuances of LGBTQ discourse for improved accuracy. Exploring sentiment shifts over time could provide insights into evolving perceptions and attitudes. Additionally, incorporating demographic information into the analysis could uncover variations in sentiment across different groups within the LGBTQ community.

Conclusion

This study's aspect-based sentiment analysis sheds light on the sentiments expressed within the LGBTQ discourse on Twitter. By comparing findings with existing literature, addressing implications, applications, limitations, and potential improvements, the study not only contributes to academic understanding but also offers insights with real-world implications. As society continues to engage in digital conversations, this research adds to the collective knowledge about LGBTQ discussions, encouraging a more informed and empathetic discourse in both online and offline spaces. It is worth mentioning that I am not a member of the LGBTQ community or interesting in anything relating to the community, but this topic was chosen to fill the gaps that were left in the Mid-Semester Hackathon and for the fact that this topic is discussed globally and has a wide range of sentiments.

Code and Dataset

https://github.com/pryyyynz/SocialMediaMining

References

1. Thakur, N. (2023). *Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox.* Atlanta: Big Data Cogn. Comput.
2. Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 74–80.
3. Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 1-167.
4. Dave, K., Lawrence, S., & Pennock, M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, (pp. 519–528).
5. Lakhani, A., Upadhyay, V., & Fiaidhi, J. (2022). Aspect Based Sentiment Analysis - Twitter. 1.
6. Katz, M., & Nandi, N. (2021). Social Media and Medical Education in the Context of the COVID-19 Pandemic: Scoping Review. *JMIR Med. Educ.*, 7.
7. Lee, H., & Cho, J. (2019). Social Media Use and Well-Being in People with Physical Disabilities: Influence of SNS and Online Community Uses on Social Support, Depression, and Psychological Disposition. *Health Commun.* , 1043–1052.
8. Kavada, A. (2015). Social Media as Conversation: A Manifesto. *Soc. Media Soc.*
9. Boon-Itt S, S. Y. (2020). Public Perception of the COVID-19 pandemic on twitter: senti- ment analysis and topic modeling study. *JMIR Public Health Surveill.*
10. Viola Savy Dsouza, P. R. (2023). A sentiment and content analysis of tweets on monkeypox stigma among the LGBTQ+ community: A cue to risk communication plan. *Dialogue in health*, 3.

11. Arcila-Calderón, C., Amores, J., P, S.-H., & Blanco-Herrero, D. (2021). Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on Twitter in Spanish. *Multimodal Technologies and Interaction*, 63.
12. Appiahene, P., Varadarajan, V., Zhang, T., & Afrifa, S. (2023). Experiences of sexual minorities on social media: A study of sentiment analysis and machine learning approaches. *Journal of Autonomous Intelligence*.
13. Jones, J. M. (2021, February 24). *Gallup*. Retrieved from LGBT Identification Rises to 5.6% in Latest U.S. Estimate: https://news.gallup.com/poll/329708/lgbt-identification-rises-latest-estimate.aspx
14. Byron, P., Rasmussen, S., Wright, T. D., Lobo, R., Robinson, K., & Paradise, B. (2017). *'You learn from each other': LGBTIQ Young People's Mental Health Help-seeking and the RAD Australia Online Directory.* Sydney: Young and Well Cooperative Research Centre.
15. Seidenberg, A. B., Jo, C., Ribisl, K., Lee, J., Butchting, F., Kim, Y., & Emery, S. (2017). *A National Study of Social Media, Television, Radio, and Internet Usage of Adults by Sexual Orientation and Smoking Status: Implications for Campaign Design.* Int. J. Environ. Res. Public Health.
16. Center, P. R. (2013, June 13). *A Survey of LGBT Americans*. Retrieved from Numbers, Facts and Trends Shaping Your World: https://www.pewresearch.org/social-trends/2013/06/13/a-survey-of-lgbt-americans/
17. MARIONETTE. (2022). *LGBT Tweets* 🏳️‍🌈🏳️‍🌈🏳️‍🌈. Retrieved from Kaggle: https://www.kaggle.com/datasets/vencerlanz09/lgbt-tweets
18. Dsouza, V. S., Rajkhowa, P., Mallya, B. R., D.S. Raksha, V. M., Cauvery, K., Raj, R., . . . Brand, H. (2023). A sentiment and content analysis of tweets on monkeypox stigma among the LGBTQ+ community: A cue to risk communication plan. *Dialogues in health*, 2.