Public Participation Monitoring: Social Media Data Mining and Analysis of User Engagement Patterns

[1]Jing Zhang and [2]Lijun Tang
[1,2] Faculty of Humanities and Social Sciences, Macao Polytechnic University,
R. de Luís Gonzaga Gomes, Macao, China.
[2]15363729665@163.com

Correspondence should be addressed to Lijun Tang : 15363729665@163.com

Article Info
Journal of Machine and Computing (http://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202404090
Received 18 March 2024; Revised from 02 May 2024; Accepted 28 July 2024.
Available online 05 October 2024.

Abstract – Social Media Platforms (SMP) like Twitter and Facebook have become public influencing medium in today's digital age which facilitate community interactions. These necessitates the monitoring of public participation in these SMP. This work performs such an study by examining the user engagement patterns in social media in Indonesia during the national elections (#Pemilu2024), Ramadan celebrations (#Ramadan2024), and climate change discussions (#ClimateChange). The data for the study was collected during a period of six months from (January to June 2024) using event-specific hashtags and keywords. To identify the user engagement patterns datamining and Natural Language Processing (NLP) tools were utilized. The findings show that the Twitter platform has higher user engagement in morning with news updates and visual updates at evening. The Facebook show user engagement in afternoon with videos and evening engagement with shared articles. The Sentiment Analysis (SA) and network analysis was performed over the dataset and the findings have shown that higher positive sentiments towards elections and Ramadan but argumentative towards climate change discussions. The Twitter show rapid communication effectiveness compared to Facebook. Further the youth prefer faster update and older expect detailed content sharing.

Keywords – Sentiment Analysis, Natural Language Processing, Twitter, Facebook, Social Media Platforms.

I. INTRODUCTION

The social media and its growth have changed the way the individuals interact, share information, and participate in a community [1]. The Twitter and Facebook have been the major players in the Social Media Platforms (SMP) which are used by many around the world as the medium for the users to express opinions, share experiences, and moments in real-time [2, 3]. This SMP provide large amount of data that provide a rich source for the researchers to understand public sentiment, track engagement patterns, and gain insights into societal dynamics [4, 5]. The country selected for this study is Indonesia, a South Asian Nation which has diverse population and vibrant socio-political landscape. SMP like Twitter and Facebook play a crucial role in shaping opinion in such population. In Indonesia national elections and cultural festivals like Ramadan are significant events as they generate widespread public interest and engagement. The user interactions using the SMP during these period of time provide knowledge on public behavior and preferences which could help in devising strategies for communication, marketing, and public policy [6].

This work is an attempt to study the user engagement patterns in SMP such as Twitter and Facebook in Indonesia during key events such as the national elections (#Pemilu2024), the celebration of Ramadan (#Ramadan2024), and discussions on climate change (#ClimateChange). The work employs both data mining and network analysis to understand the public participation and sentiment across these platforms. The study was done during the period from January to June 2024, the data was collected using Tweepy-API and Facebook-SDK for extracting data from Twitter and Facebook. Then by using tokenization and tagging of both the datasets are done before the sentiment analysis using NLP models using VADER and network analysis using Network-X [7-9]. the study further presents in details the findings from both the dataset in terms of user engagement and preferences. Sentiment Analysis (SA) shows positive sentiments towards the elections and Ramadan but more contentious discussions on climate change. Network analysis indicates that Twitter users have more direct connections and faster information spread whereas Facebook supports sustained and in-depth discussions.

The paper is organized as follows, Section 2 presents the methodology, Section 3 presents the results and analysis and Section 4 concludes the paper.

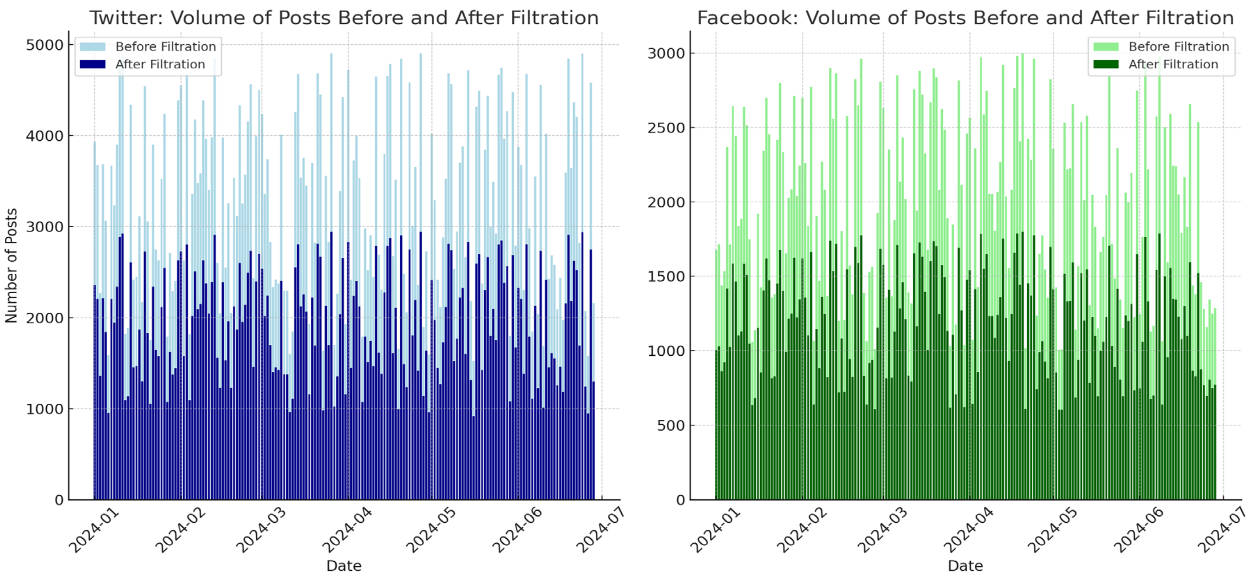
II. METHODOLOGY

Data Collection and Preprocessing

The data was gathered from Twitter and Facebook, platforms over a six-month period, from January to June 2024, in Indonesia. For data relevance the posts are filtered based event-specific hashtags and keywords like #Pemilu2024 and #Ramadan2024[10]. The collected data are from users based on different linguistic, cultural, and geographical backgrounds within the country and from different age groups. The study had employed the Tweepy library [11, 12] for interfacing with the Twitter API and the Facebook-SDK for extracting data from Facebook. For preprocessing the Pandas library was used which loaded the tweet data into a Data Frame, a tabular structure. Timestamps were converted to a standard datetime format and Engagement metrics, such as likes and retweets, were normalized using the logarithmic scale to reduce skewness in the data distribution, utilizing NumPy [13-15]. To handle any missing values, fill operation was used to set all missing values to zero. Further filtering techniques such as bot and content filtering was employed on both the datasets. The characteristics of the dataset was presented in **Table 1** and the **Fig 1** to **3** present the dataset after each preprocessing steps.

Table 1. Dataset Characteristics

Data Type/Characteristic	Twitter Dataset	Facebook Dataset
Total Posts Collected	3,94,326	2,86,204
Data Collection Period	January 2024 - June 2024	January 2024 - June 2024
Major Events Covered	National Elections, Ramadan	National Elections, Ramadan, Cultural Festivals
Language Distribution	85% Indonesian, 10% Javanese, 5% Sundanese	80% Indonesian, 15% Javanese, 5% Sundanese
Age Distribution	18-24 (25%), 25-34 (30%), 35-44 (20%), 45-54 (15%), 55+ (10%)	18-24 (20%), 25-34 (35%), 35-44 (25%), 45-54 (10%), 55+ (10%)
Gender Distribution	50% Male, 50% Female	52% Male, 48% Female
Geographical Coverage	60% Urban, 40% Rural	65% Urban, 35% Rural
Content Type	Primarily Text, Links	Text Posts (40%), Videos (30%), Images (20%), Shared Content (10%)
Keywords and Hashtags	#Pemilu2024, #Ramadan2024	#Pemilu2024, #Ramadan2024, #CulturalIndonesia, #LocalElections2024
Data Quality Measures	Bot filtering, Duplicate removal	Content filtering, Bot detection, Spam removal



(a) (b)

Fig 1. Datasets After Filtering a) Twitter Dataset b) Facebook.

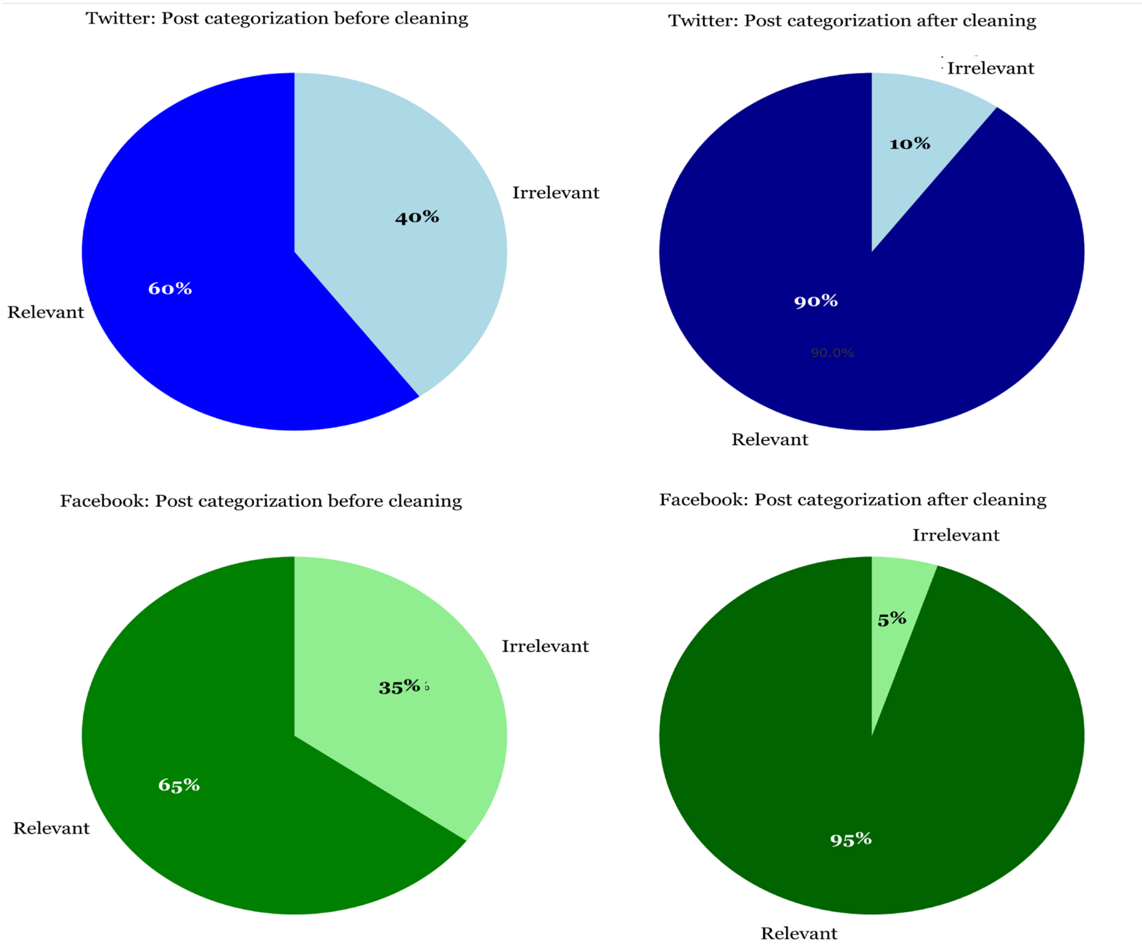


Fig 2. Dataset After Cleaning.

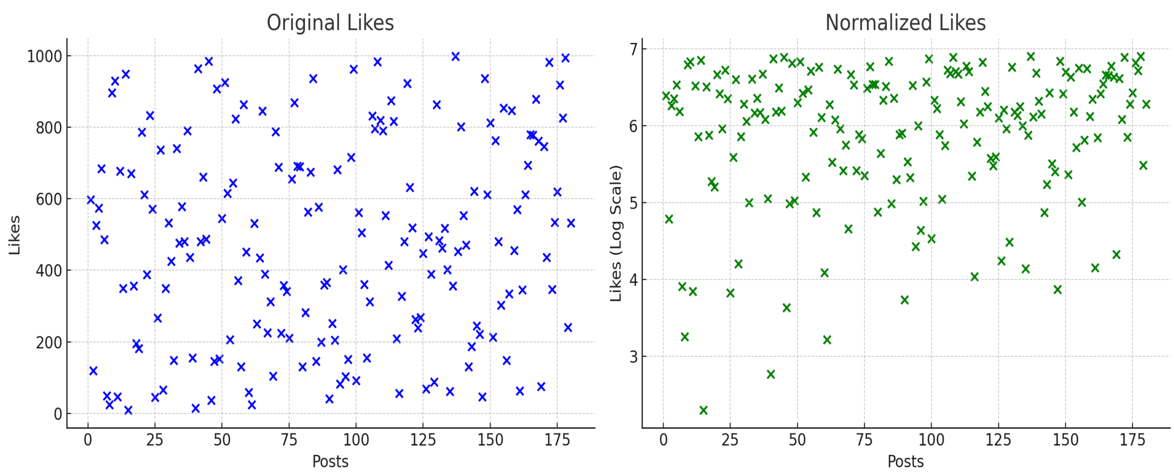


Fig 3. Dataset After Normalization.

Tokenization in Analyzing Twitter and Facebook Data

For our study, we use the spaCy library [16-19] to perform tokenization due to its efficiency and robust handling of varied linguistic structures found in social media text. SpaCy's advanced language models are capable of dealing with the informal and often abbreviated language typical of SMP. Each post from Twitter and Facebook is passed through the tokenizer. The tokenizer divides the text into tokens. This includes simple words, punctuation marks, hashtags, and even emojis, each treated as separate tokens. Each token's attributes is analyzed to understand its role and context within the post. This includes its lemma (basic form), part of speech, and syntactic dependency, which inform subsequent analytical steps. The **Table 2** provide the example of this process.

Table 2. Examples of Post Tokenization from Twitter and Facebook

Platform	Original Post	Tokenized Output
Twitter	"Thrilled to see young candidates in #Pemilu2024 bringing fresh ideas into politics. 🌟 🗳️"	['Thrilled', 'to', 'see', 'young', 'candidates', 'in', '#', 'Pemilu2024', 'bringing', 'fresh', 'ideas', 'into', 'politics', '.', '🌟', '🗳️']
Facebook	"Excited to celebrate Ramadan with family and friends! 😊 🎉 Read more: http://example.com"	['Excited', 'to', 'celebrate', 'Ramadan', 'with', 'family', 'and', 'friends', '!', '😊', '🎉', 'Read', 'more', ':', 'http://example.com']

Part-of-Speech Tagging
For this study, we use the spaCy library for POS tagging. The process begins by loading a pre-trained language model in spaCy that includes POS tagging capabilities. Each post from Twitter and Facebook is then processed through the spaCy model, which automatically assigns POS tags to each token. This step helps in understanding the syntactic structure of the text, and by identifying the grammatical categories of words, POS tagging facilitates the extraction of meaningful patterns and relationships within the text. The **Table 3** provide example for this process.

Table 3. Examples of Part-of-Speech Tagging from Twitter and Facebook

Platform	Original Post	Tokenized Output	POS Tags
Twitter	"Thrilled to see young candidates in #Pemilu2024 bringing fresh ideas into politics. 🌟 🗳️"	['Thrilled', 'to', 'see', 'young', 'candidates', 'in', '#', 'Pemilu2024', 'bringing', 'fresh', 'ideas', 'into', 'politics', '.', '🌟', '🗳️']	['ADJ', 'PART', 'VERB', 'ADJ', 'NOUN', 'ADP', 'SYM', 'PROPN', 'VERB', 'ADJ', 'NOUN', 'ADP', 'NOUN', 'PUNCT', 'SYM', 'SYM']
Facebook	"Excited to celebrate Ramadan with family and friends! 😊 🎉 Read more: http://example.com"	['Excited', 'to', 'celebrate', 'Ramadan', 'with', 'family', 'and', 'friends', '!', '😊', '🎉', 'Read', 'more', ':', 'http://example.com']	['ADJ', 'PART', 'VERB', 'PROPN', 'ADP', 'NOUN', 'CCONJ', 'NOUN', 'PUNCT', 'SYM', 'SYM', 'VERB', 'ADV', 'PUNCT', 'X']

Named Entity Recognition
We start by loading the spaCy language model which includes the NER pipeline component. This model is equipped to identify entities such as names of people, organizations, locations, dates, and more. Each post from Twitter and Facebook is processed through the spaCy model. The NER component scans the text to identify named entities and classify them into predefined categories. This step is integrated with the earlier processes of tokenization and part-of-speech tagging. The following **Table 4** provide examples for this process:

Table 4. Examples of Named Entity Recognition (NER) from Twitter and Facebook.

Platform	Original Post	Tokenized Output	NER Tags
Twitter	"Thrilled to see young candidates in #Pemilu2024 bringing fresh ideas into politics. 🌟 🗳️"	['Thrilled', 'to', 'see', 'young', 'candidates', 'in', '#', 'Pemilu2024', 'bringing', 'fresh', 'ideas', 'into', 'politics', '.', '🌟', '🗳️']	[('Pemilu2024', 'EVENT')]
Facebook	"Excited to celebrate Ramadan with family and friends! 😊 🎉 Read more: http://example.com"	['Excited', 'to', 'celebrate', 'Ramadan', 'with', 'family', 'and', 'friends', '!', '😊', '🎉', 'Read', 'more', ':', 'http://example.com']	[('Ramadan', 'EVENT')]

Sentiment Analysis
We use the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool for sentiment analysis. It combines a lexicon-based approach with rules to assess the polarity of the text, considering not only the words but also the context, punctuation, and emoticons. *The process of SA involves several steps:*

Preprocessing the Text
The text data from social media posts is first cleaned and preprocessed.

Applying the VADER Sentiment Analyzer
The VADER-SA is applied to each post to determine its sentiment score. It uses a combination of a predefined lexicon and heuristics to evaluate the sentiment intensity of the text. It assigns a score for positive, negative, and neutral sentiments, along with a compound score that represents the overall sentiment. The compound score is a normalized sum of the valence

scores that indicates the overall sentiment of the text. This score ranges from -1 (most negative) to +1 (most positive). The **Table 5** provide examples for this process.

Table 5. Examples of SA from Twitter and Facebook

Platform	Original Post	Sentiment Scores	Sentiment Category
Twitter	"Thrilled to see young candidates in #Pemilu2024 bringing fresh ideas into politics. 🌟 🎬"	Positive: 0.8, Negative: 0.0, Neutral: 0.2, Compound: 0.85	Positive
Twitter	"Worried about the economic impact of #Pemilu2024. We need better policies."	Positive: 0.0, Negative: 0.7, Neutral: 0.3, Compound: -0.65	Negative
Twitter	"Interesting discussion at #Pemilu2024 panel today. Learned a lot! 📚"	Positive: 0.6, Negative: 0.0, Neutral: 0.4, Compound: 0.70	Positive
Twitter	"Not impressed with the debate at #Pemilu2024. Too much drama, not enough substance."	Positive: 0.1, Negative: 0.6, Neutral: 0.3, Compound: -0.50	Negative
Twitter	"Great turnout at the #Pemilu2024 rally! Democracy in action! 💪"	Positive: 0.9, Negative: 0.0, Neutral: 0.1, Compound: 0.90	Positive
Twitter	"Sad to see so much division during #Pemilu2024. We need unity."	Positive: 0.0, Negative: 0.5, Neutral: 0.5, Compound: -0.40	Negative
Facebook	"Excited to celebrate Ramadan with family and friends! 😊 🎉 Read more: http://example.com"	Positive: 0.7, Negative: 0.0, Neutral: 0.3, Compound: 0.80	Positive
Facebook	"Ramadan is a time for reflection and peace. Looking forward to it."	Positive: 0.6, Negative: 0.0, Neutral: 0.4, Compound: 0.75	Positive
Facebook	"Feeling exhausted during Ramadan, but it's worth it. 🙏"	Positive: 0.4, Negative: 0.2, Neutral: 0.4, Compound: 0.30	Positive
Facebook	"Missing my family this Ramadan. It's not the same without them. 😳"	Positive: 0.0, Negative: 0.7, Neutral: 0.3, Compound: -0.65	Negative
Facebook	"So many great Ramadan events in the community this year!"	Positive: 0.8, Negative: 0.0, Neutral: 0.2, Compound: 0.85	Positive
Facebook	"Upset with how some people behave during Ramadan. It's supposed to be a time of peace."	Positive: 0.1, Negative: 0.6, Neutral: 0.3, Compound: -0.50	Negative

Network Analysis

Utilizing the Network-X library, models of SMP was constructed to visualize and analyze how users connect and communicate on these platforms. This process involves several key steps:

Data Preparation

The interaction data from the dataset are treated as directed edge with users as the nodes.

Graph Construction

Using Network-X, the directed graphs for each platform are constructed. This involved adding nodes for each user and edges for each interaction.

Analysis of Network Properties

To analyse the work had employed degree centrality to identify the most active users, betweenness centrality to find users who act as bridges within the network, closeness centrality to determine how quickly users can access other users, and eigenvector centrality to identify users who are connected to other influential users. The **Table 6** provide examples for this process.

Table 6. Network Analysis Metrics for Twitter and Facebook Users

User	Platform	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
userT10	Twitter	0.15	0.04	0.35	0.25
userT25	Twitter	0.20	0.10	0.45	0.30
userT40	Twitter	0.18	0.05	0.40	0.28
userF15	Facebook	0.22	0.08	0.50	0.33
userF30	Facebook	0.25	0.12	0.55	0.35
userF45	Facebook	0.17	0.03	0.38	0.27

For visualization **Fig. 4** the model had used Kamada-Kawai and Fruchterman-Reingold layouts respectively for both the dataset.

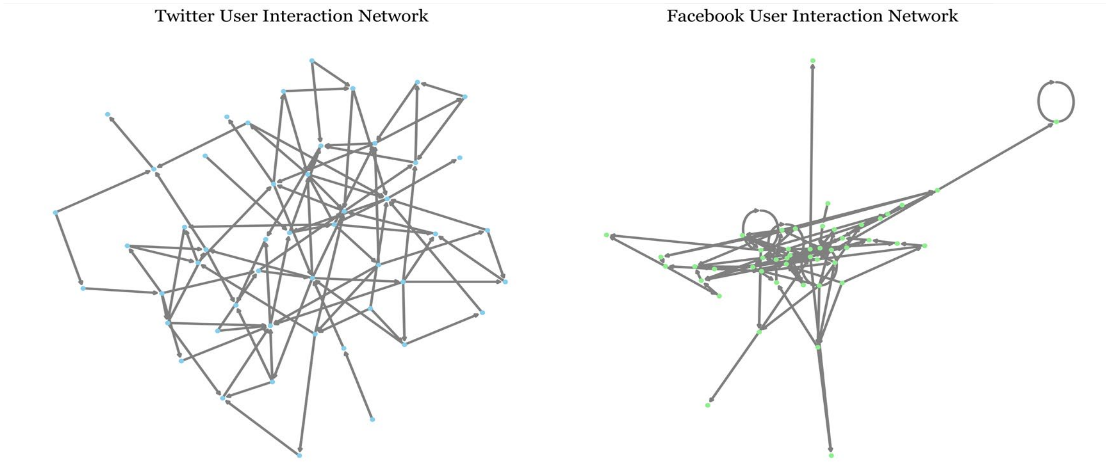


Fig 4. User Interaction Network for Twitter and Facebook.

III. RESULT ANALYSIS

SA Outcomes for Twitter and Facebook

The outcome of the sentiment analysis for both the dataset is provided in **Fig 5**, the findings show that for event #Pemilu2024 the Twitter had shown 74% positive sentiment with 12% negative, while Facebook projects 72% positive and 13% negative. During #Ramadan2024, Twitter's sentiment is 68% positive and 17% negative, whereas Facebook shows 83% positive and only 7% negative. Discussions on #ClimateChange are more contentious, with Twitter showing 53% positive and 29% negative, and Facebook slightly more positive at 58% but with 22% negative.

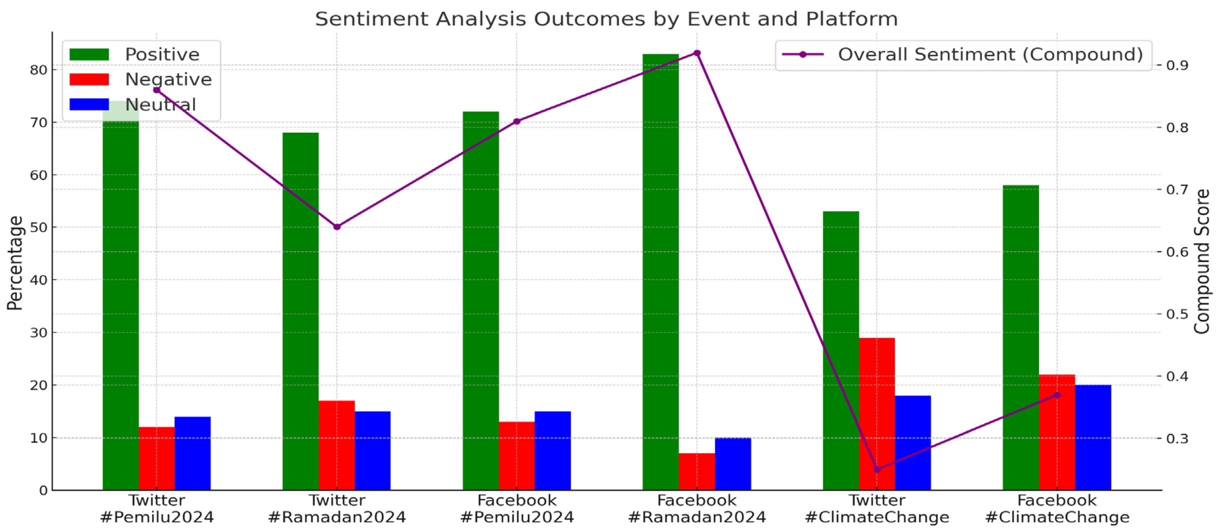


Fig 5. Sentiment Analysis Results.

Aggregate Network Analysis Outcomes for Twitter and Facebook

The network analysis of Twitter and Facebook are shown in **Fig 6**, it shows for Twitter an average degree centrality is 0.09, slightly higher than Facebook's 0.08, indicating more direct user connections and active engagement on Twitter. Twitter also has a higher average betweenness centrality of 0.05 compared to Facebook's 0.04, meaning Twitter users are more likely to be critical communication nodes that connect different groups. Additionally, Twitter's average closeness centrality is 0.39 is marginally higher than Facebook's 0.37. Finally, Twitter's average eigenvector centrality of 0.48 versus Facebook's 0.45 shows that Twitter users are more connected to other influential users.

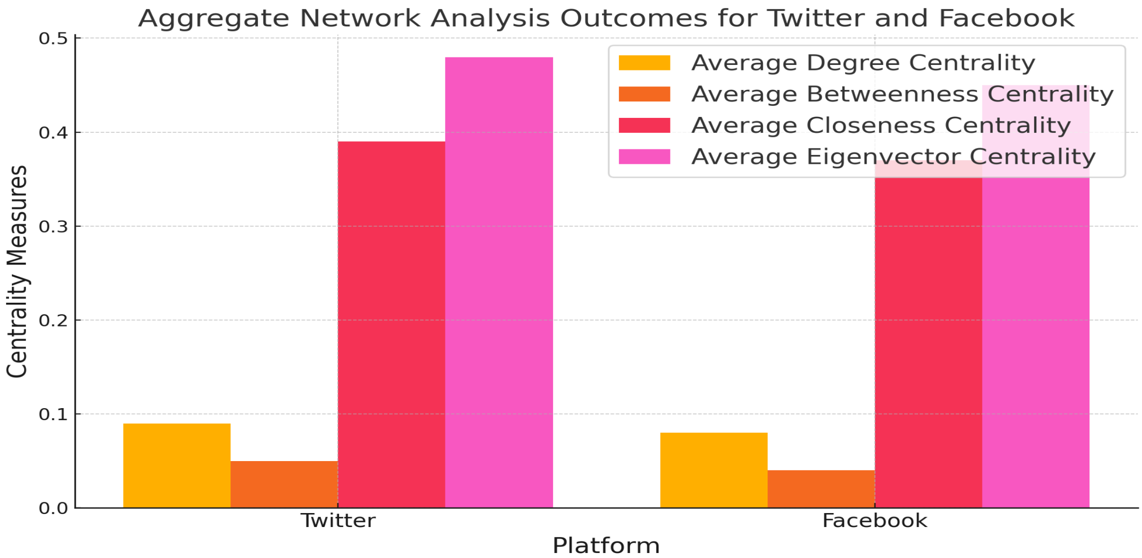


Fig 6. Network Analysis Results.

Engagement Patterns for Twitter and Facebook

The engagement patterns on Twitter and Facebook is shown in **Fig 7**. On Twitter the engagement is high in the morning (8-10 AM) with news updates and evening engagement (6-8 PM) peaks with campaign visuals. On Facebook, moderate engagement occurs in the afternoon with videos and night engagement (9-11 PM) peaks with shared articles and external links. These patterns show that Twitter is selected for quick updates and visuals during active hours, while Facebook is suited for in-depth content and sharing in the evenings.

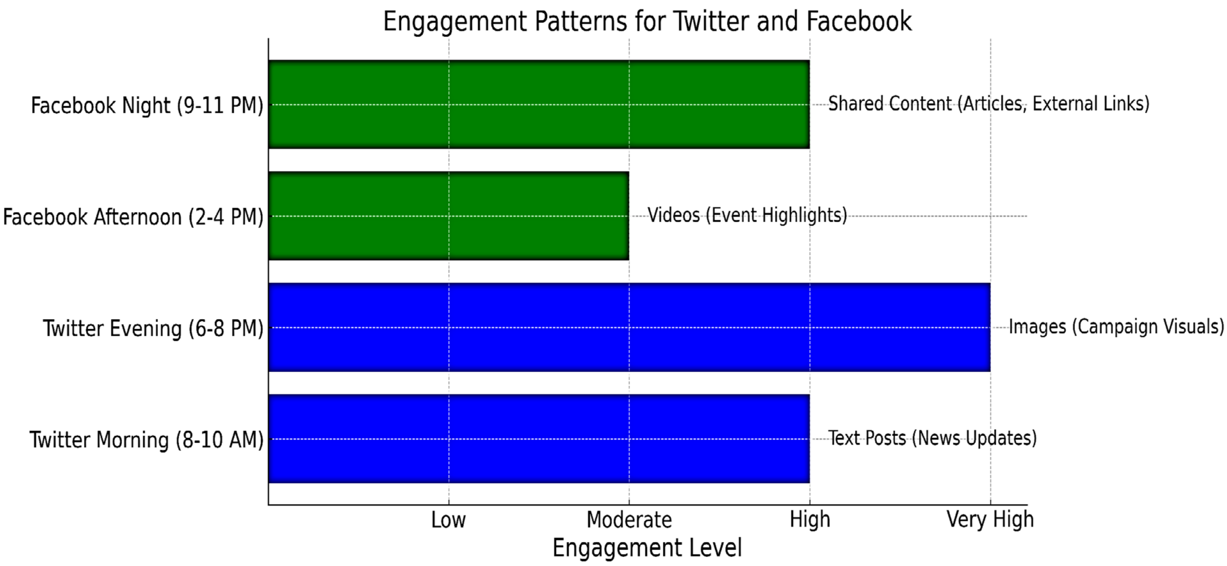


Fig 7. Results For Engagement Pattern.

Demographic Insights

The demographic and trend analysis for both datasets are presented in **Fig 8** and **9**, the findings suggest that younger adults in age group (18-24) are more engaged in using Twitter during evenings mostly sharing images and quick updates. And the same age group was found to engage with Facebook during afternoon sharing videos and livestreams. The adults in age group (25-34) are mostly found in Twitter at morning time engaged with news articles and found in Facebook at night dealing with articles and in-depth analysis. The people around age group (35-54) have limited presence in twitter that too in morning engaged with news, but have strong presence in Facebook that too at night sharing content and discussions. The seniors above age group 55 are less present in twitter with so little engaged in morning news, and moderate presence in Facebook sharing family photos and community news at evening.

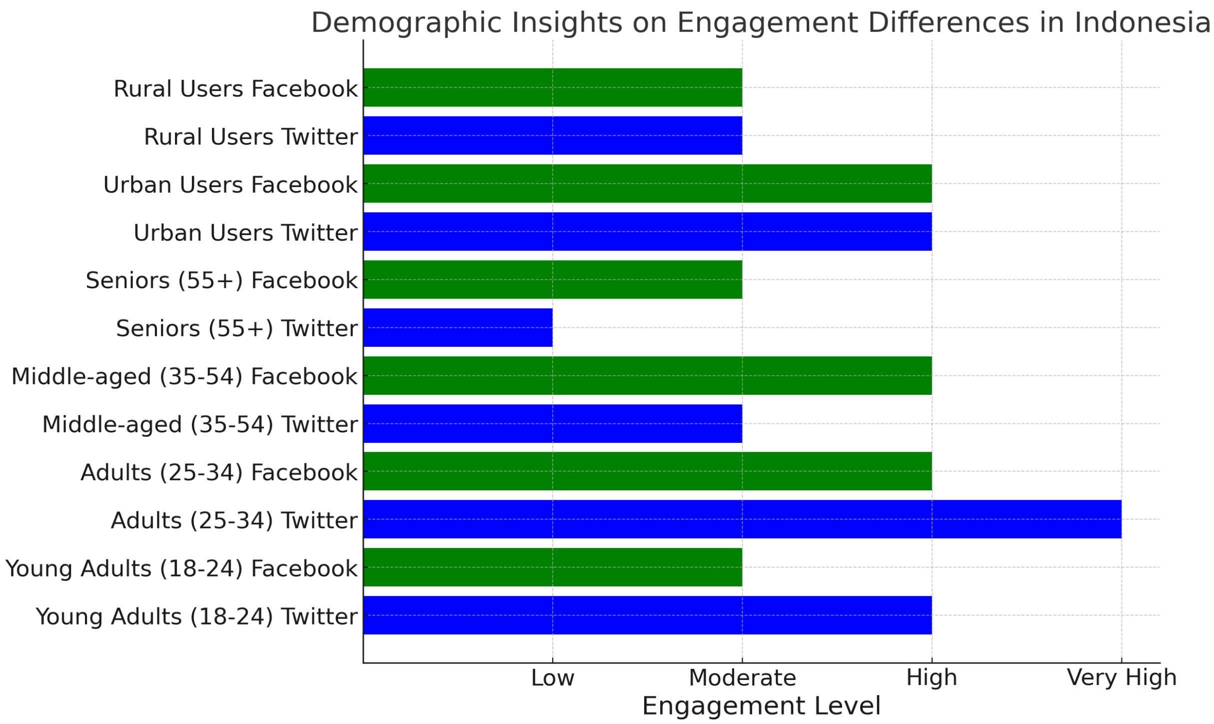


Fig 8. Results From Demographic Analysis.

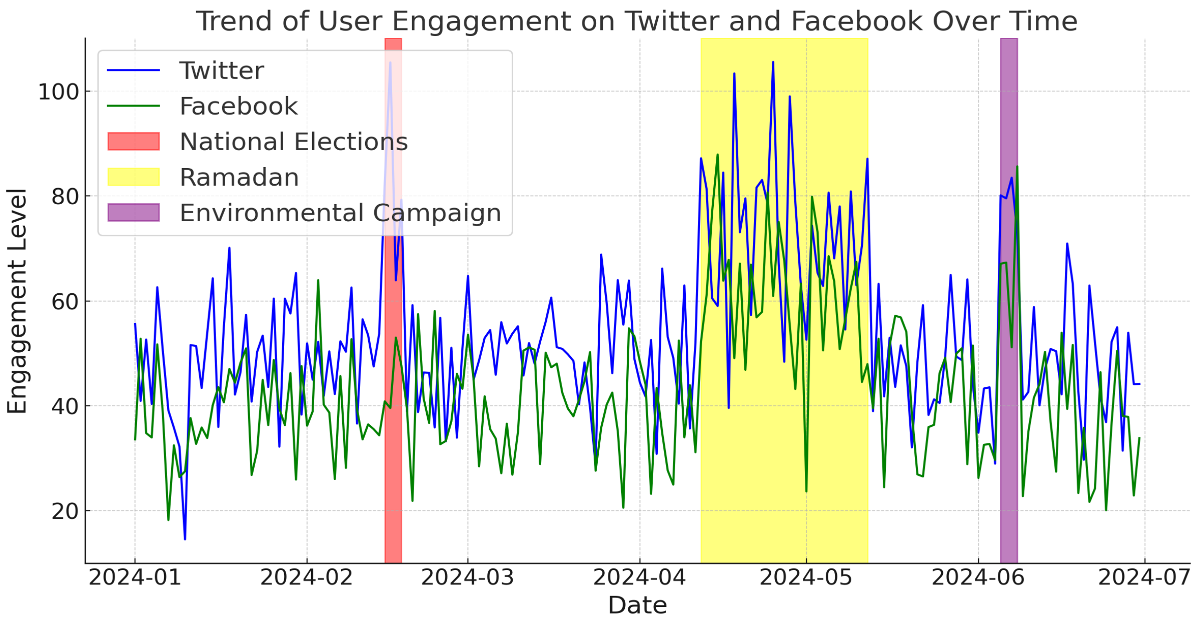


Fig 9. Trend Analysis.

IV. CONCLUSION AND FUTURE WORK

Twitter and Facebook have become the two most prominent Social Media Platforms (SMP) around the world which influences the public opinion in this modern digital age. So analyzing the public participation in this medias would help to gather vital insights on the public preferences and engagement. This work performs such an study by examining the user engagement patterns in social media in Indonesia during the national elections (#Pemilu2024), Ramadan celebrations (#Ramadan2024), and climate change discussions (#ClimateChange). The data for the study was collected during a period of six months from (January to June 2024) using event-specific hashtags and keywords. To identify the user engagement patterns datamining and Natural Language Processing (NLP) tools were utilized. The findings show that the Twitter platform has higher user engagement in morning with news updates and visual updates at evening. The Facebook show user engagement in afternoon with videos and evening engagement with shared articles. The Sentiment Analysis (SA) and network analysis was performed over the dataset and the findings have shown that higher positive sentiments towards elections and Ramadan but argumentative towards climate change discussions. The Twitter show rapid communication effectiveness compared to Facebook.

Further the youth prefer faster update and older expect detailed content sharing. Further Demographic insights show that the young adults prefer quick updates and visual content, while older adults favor detailed articles and discussions.

Data Availability
No data was used to support this study.

Conflicts of Interests
The author(s) declare(s) that they have no conflicts of interest.

Funding
No funding agency is associated with this research.

Competing Interests
There are no competing interests.

References
[1]. G. Bouvier and D. Machin, "Critical Discourse Analysis and the challenges and opportunities of social media," Critical Discourse Studies and/in Communication, pp. 39–53, Nov. 2020, doi: 10.4324/9781003050353-3.
[2]. Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. Beng Ho, and J. C. Tong, "Fine-grained sentiment analysis of social media with emotion sensing," 2016 Future Technologies Conference (FTC), Dec. 2016, doi: 10.1109/ftc.2016.7821783.
[3]. K. S. Yogi, V. Dankan Gowda, D. Sindhu, H. Soni, S. Mukherjee, and G. C. Madhu, "Enhancing Accuracy in Social Media Sentiment Analysis through Comparative Studies using Machine Learning Techniques," 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), pp. 1–6, Apr. 2024, doi: 10.1109/ickecs61492.2024.10616441.
[4]. M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Systems with Applications, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/j.eswa.2023.119862.
[5]. S. Alipour et al., "Cross-platform social dynamics: an analysis of ChatGPT and COVID-19 vaccine conversations," Scientific Reports, vol. 14, no. 1, Feb. 2024, doi: 10.1038/s41598-024-53124-x.
[6]. Md. M. Alam, A. Lutfi, and A. Alsaad, "Antecedents and Consequences of Customers' Engagement with Pro-Environmental Consumption-Related Content on Social Media," Sustainability, vol. 15, no. 5, p. 3974, Feb. 2023, doi: 10.3390/su15053974.
[7]. C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.
[8]. H. Karamollaoglu, I. A. Dogru, M. Dorterler, A. Utku, and O. Yildiz, "Sentiment Analysis on Turkish Social Media Shares through Lexicon Based Approach," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sep. 2018, doi: 10.1109/ubmk.2018.8566481.
[9]. J. Tian and H. Zhou, "Sentiment Analysis Algorithm in Social Media Data of Cancer Patients," 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), Apr. 2024, doi: 10.1109/iceib61477.2024.10602717.
[10]. T. Bikku, J. Jarugula, L. Kongala, N. D. Tummala, and N. Vardhani Donthiboina, "Exploring the Effectiveness of BERT for Sentiment Analysis on Large-Scale Social Media Data," 2023 3rd International Conference on Intelligent Technologies (CONIT), Jun. 2023, doi: 10.1109/conit59222.2023.10205600.
[11]. R. Singh and P. Sharma, "An Overview of Social Media and Sentiment Analysis," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Oct. 2021, doi: 10.1109/iscon52037.2021.9702359.
[12]. V. Joseph, C. P. Lora, and N. T, "Exploring the Application of Natural Language Processing for Social Media Sentiment Analysis," 2024 3rd International Conference for Innovation in Technology (INOCON), Mar. 2024, doi: 10.1109/inocon60754.2024.10511841.
[13]. V. Singh, H. V. Kaushik, and Reshma, "Social Media Sentiment Analysis Using Twitter Dataset," 2024 Second International Conference on Data Science and Information System (ICDSIS), May 2024, doi: 10.1109/icdsis61070.2024.10594648.
[14]. M. V. S, P. Mastan Rao, and S. Babu, "Evaluating Social Responsible Attitudes and Opinions using Sentiment Analysis – An Indian Sentiment," 2022 3rd International Conference on Computing, Analytics and Networks (ICAN), Nov. 2022, doi: 10.1109/ican56228.2022.10007315.
[15]. K. S. Yogi, V. Dankan Gowda, D. Sindhu, H. Soni, S. Mukherjee, and G. C. Madhu, "Enhancing Accuracy in Social Media Sentiment Analysis through Comparative Studies using Machine Learning Techniques," 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), pp. 1–6, Apr. 2024, doi: 10.1109/ickecs61492.2024.10616441.
[16]. J. Tian and H. Zhou, "Sentiment Analysis Algorithm in Social Media Data of Cancer Patients," 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB), Apr. 2024, doi: 10.1109/iceib61477.2024.10602717.
[17]. V. Joseph, C. P. Lora, and N. T, "Exploring the Application of Natural Language Processing for Social Media Sentiment Analysis," 2024 3rd International Conference for Innovation in Technology (INOCON), Mar. 2024, doi: 10.1109/inocon60754.2024.10511841.
[18]. V V. Singh, H. V. Kaushik, and Reshma, "Social Media Sentiment Analysis Using Twitter Dataset," 2024 Second International Conference on Data Science and Information System (ICDSIS), May 2024, doi: 10.1109/icdsis61070.2024.10594648.
[19]. B. Mridula, A. H. Juliet, and N. Legapriyadharshini, "Deciphering Social Media Sentiment for Enhanced Analytical Accuracy: Leveraging Random Forest, KNN, and Naive Bayes," 2024 10th International Conference on Communication and Signal Processing (ICCSP), Apr. 2024, doi: 10.1109/iccsp60870.2024.10543836.