Sentiment Analysis of Twitter Data on the 2024 Indonesian Presidential Election Using BERT

Ahmad Roihan[1], Tito Tri Atmojo[*2], Rizky A Wardoyo[3], Muhamad S T Saputra[4]
[1,3,4] Department of Computer Systems, University of Raharja, Indonesia
[2] Informatics Engineering Department, University of Raharja, Indonesia
E-mail: [1]ahmad.roihan@raharja.info, [*2]**tito.tri@raharja.info**, [3]rizky.adytya@raharja.info, [4]stabil@raharja.info

Abstract

Social media platforms, particularly Twitter, are frequently employed by individuals to articulate their opinions on various subjects in textual form. The proliferation of viewpoints from diverse sources can influence public perceptions on these topics. The greater the popularity of a topic, the more abundant the opinions generated. Currently, the most widely discussed topic is the 2024 Indonesian presidential election. Sentiment analysis, or opinion mining, is an academic discipline that examines sentiments towards a given entity, while text mining involves the extraction of information through processing, classifying, and analyzing extensive datasets. This study will utilize data crawling techniques to gather data from Twitter which will subsequently undergo preprocessing and cleaning. Following this, the cleaned data will be classified by sentiment (positive, negative, or neutral) using a pre-trained language model (BERT) and Natural Language Toolkit (NLTK). The classified data will then be visualized with tools such as Matplotlib and Wordcloud to elucidate the data distribution.

Keywords *— Social Media Platforms, Indonesian Presidential Election 2024, BERT*

1. INTRODUCTION

In the contemporary digital era, a significant portion of individuals depend on social media for information acquisition. Among the myriad social media platforms, Twitter stands out as a prevalently utilized medium where users disseminate their viewpoints through textual expressions and engage in global discourse. The topics most frequently deliberated by the public on Twitter are prominently featured as trending topics. Twitter endows users with the platform to articulate their opinions, emotions, and perspectives. This social medium harbors an extensive repository of information encapsulated in tweets.

The 2024 Indonesian Presidential Election transpired in Indonesia on February 14, 2024. This juncture presents a propitious moment to scrutinize Indonesian public sentiment towards the election, given the substantial volume of information and viewpoints conveyed by the populace. Considering the sheer expanse of data encapsulated in tweets [1][2], it is imperative to conduct an analysis to comprehend the public's reactions to the election [3].

This paper delves into sentiment analysis [4][5] with the objective of elucidating public sentiment regarding the 2024 Indonesian Presidential Election by aggregating tweets, which encompass a diverse array of random and unstructured texts. These tweets typically exhibit sentiments that are positive, negative, or neutral [6]. By conducting sentiment analysis on

Indonesian public opinions, the study aims to yield valuable insights into the populace's perspectives on the election. For this research, data pertinent to the topic has been collected through a data crawling methodology, resulting in a csv file comprising 597 tweets

2. RESEARCH METHOD

In this study, data pertaining to the topic was amassed in text form using a data crawling technique, resulting in a CSV file that encompassed 597 tweets. The file includes various tweets containing textual content, hashtags, links/URLs, punctuation, special characters, emoticons, and user mentions. Initially, the data undergoes preprocessing through Natural Language Processing (NLP) to remove extraneous elements such as hashtags, links/URLs, punctuation, special characters, emoticons, and user mentions. Subsequently, the sentiment of the tweets—categorized as positive, negative, or neutral—is determined utilizing a pre-trained language model (BERT) [7].

Figure 1. Research method

2.1. Data Collection

During the data crawling procedure, tweets pertinent to the 2024 Indonesian presidential election will be gathered from Twitter over the period spanning January 01, 2024, to March 25, 2024. The tweets will be acquired using a Twitter crawler named Tweet Harvest, and subsequently stored in a CSV file format. Data collection or data crawling is the process of retrieving data by scrapping social media [8].

2.2. Data Processing

The accumulated data will undergo a pre-processing stage where extraneous elements will be removed from the raw data in each tweet. This data processing step aims to organize the data structure, making it more streamlined and easier to classify. In the field of sentiment analysis, pre-processing techniques are used to rework the text so that it can be better understood by classification systems, for example by reducing noise and reorganizing the content. As a result, along with the development of methods and networks, over the years several pre-processing mechanisms have been progressively implemented and tested by researchers [9].

2.3. Data Classification

The data, once collected and cleansed, will be categorized into negative, positive, or neutral sentiments. Following classification, the data will be visualized in a graph to illustrate the distribution of each sentiment.

2.4. Evaluation

Subsequent to the classification process, an analysis will be conducted to compare the quantity of negative, positive, and neutral sentiments presented in the graph, leading to the study's conclusions.

3. RESEARCH RESULTS AND DISCUSSION

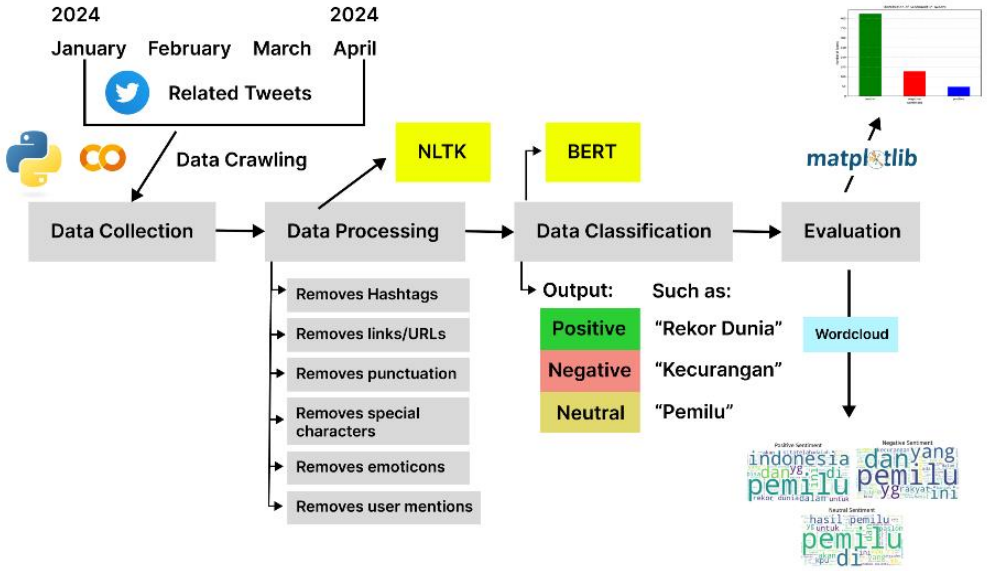

Figure 2. Sentiment Analysis Workflow with Tools

3.1. Data Collection

The text data utilized in this study was sourced from social media platform Twitter, encompassing 597 tweets containing the keyword “pemilu 2024” within the period from January 1, 2024, to March 25, 2024.

	conversation_id_str	created_at	favorite_count	full_text	id_str
0	1.770000e+18	Sun Mar 17 10:43:11 +0000 2024	333	Jika KPU (@KPU_ID) profesional tentu semua kea...	1.770000e+18

Figure 3. One Data Sample

3.2. Data Processing

The gathered data will undergo a preprocessing phase to cleanse each tweet of extraneous elements such as hashtags, links/URLs, punctuation, special characters, emoticons, and user mentions. This process will utilize NLTK in Python for data preprocessing.

Table 1. The Original Text and Preprocessed Text

Original Text	Preprocessed Text
Jika KPU (@KPU_ID) profesional tentu semua keanehan termasuk soal temuan kecurangan pemilu di luar negeri tidak akan pernah terjadi. Ikuti pembahasannya di #BedahEditorial di kanal youtube metrotvnews https://t.co/TMaoS98MMV #Menunggu #Tanggungjawab #KPU #Pemilu2024 https://t.co/kqC5oTnBHI	jika kpu profesional tentu semua keanehan termasuk soal temuan kecurangan pemilu di luar negeri tidak akan pernah terjadi ikuti pembahasannya di di kanal youtube metrotvnews tcotmaos98mmv tcokqc5otnbhi
Unjuk rasa terjadi di luar gedung KPU RI menuntut pasangan capres-cawapres Prabowo-Gibran untuk didiskualifikasi (18/3). Massa mengajukan tuntutan karena Pemilu 2024 dinilai sarat dengan kecurangan. https://t.co/fm1hOSBRcM	unjuk rasa terjadi di luar gedung kpu ri menuntut pasangan caprescawapres prabowogibran untuk didiskualifikasi 183 massa mengajukan tuntutan karena pemilu 2024 dinilai sarat dengan kecurangan tcofm1hosbrcm
TERUS BERJUANG TETAP BERSATU â€¼ï¸• THN AMIN: Proses MK jalan HAK ANGKET JALAN Proses hukum jalan Proses politik jalan Gerakan masyarakan jalan LAWAN KECURANGAN PEMILU 2024 â€¼ï¸• https://t.co/9zGZKxE2sU	terus berjuang tetap bersatu thn amin proses mk jalan hak angket jalan proses hukum jalan proses politik jalan gerakan masyarakan jalan lawan kecurangan pemilu 2024 tco9zgzkxe2su
Waktu Data Masuk 0 00% Paslon 02 sudah 57 69% . Masih Percaya Pilpres Nggak Curang??? Tgl 13 Pebruari 2024 Lupa robah Tanggal?? Padahal Pemilu tgl 14 Pebruari 2024 https://t.co/YVxlA8lHRU	waktu data masuk 0 00 paslon 02 sudah 57 69 masih percaya pilpres nggak curang tgl 13 pebruari 2024 lupa robah tanggal padahal pemilu tgl 14 pebruari 2024 tcoyvxla8lhru
Kami tidak akan membiarkan penyimpangan atas demokrasi ini berlalu tanpa catatan kami tidak ingin ini menjadi preseden yang buruk bagi generasi - generasi mendatang. Biar cukup berhenti sampai disini #1PKB #Pemilu2024 https://t.co/6EzUXSlIfY	kami tidak akan membiarkan penyimpangan atas demokrasi ini berlalu tanpa catatan kami tidak ingin ini menjadi preseden yang buruk bagi generasi generasi mendatang biar cukup berhenti sampai disini tco6ezuxslify

Subsequently, the data will be categorized by sentiment using a pre-trained language model (BERT), as illustrated in Figure 3, which demonstrates the sentiment classification of the processed texts.

	preprocessed_text	sentiment
0	jika kpu profesional tentu semua keanehan ter...	neutral
1	unjuk rasa terjadi di luar gedung kpu ri menun...	neutral
2	terus berjuang tetap bersatu thn amin proses ...	neutral
3	waktu data masuk 0 00 paslon 02 sudah 57 69 m...	negative
4	kami tidak akan membiarkan penyimpangan atas d...	negative

Figure 4. The Preprocessed Texts

3.3. Data Classification

The sentiment analysis results of the 597 tweets concerning the 2024 Indonesian presidential election are presented. The classified data will then be visualized with tools such as Matplotlib and Wordcloud to elucidate the data distribution. Matplotlib is used to create a graph representing the data distribution below.

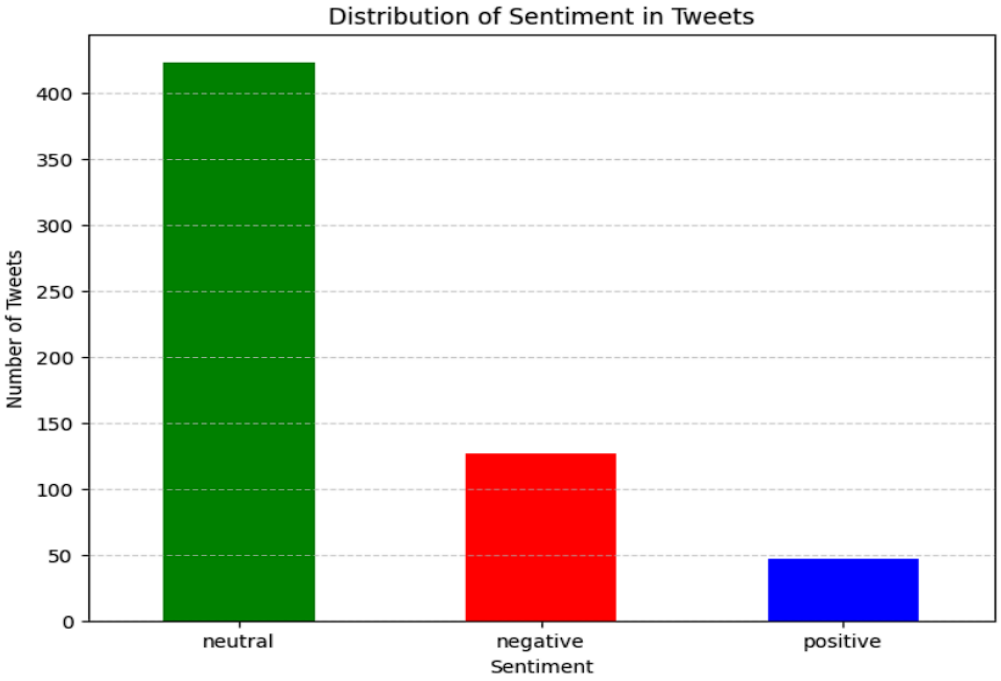

Figure 5. The Distribution of Sentiment

3.4. Evaluation

Figure 4 categorizes the sentiment distribution of the tweets: 423 tweets exhibit neutral sentiment, 127 tweets show negative sentiment, and 47 tweets express positive sentiment. Figure 5 provides a visual representation of the sentiment analysis outcomes.

```
sentiment
neutral     423
negative    127
positive     47
Name: count, dtype: int64
```

Figure 6. The Number of Tweets

Key terms associated with the positive sentiment include “rekor dunia” and “terimakasih” which convey a favorable perception of the 2024 Indonesian presidential election. Conversely, terms such as “kecurangan” and “bansos” prominently feature in the negative sentiment category, indicating a critical view of the election. WordCloud will then be used to generate the graph below, illustrating the most frequent words within each sentiment category, with word size indicating the frequency of each word

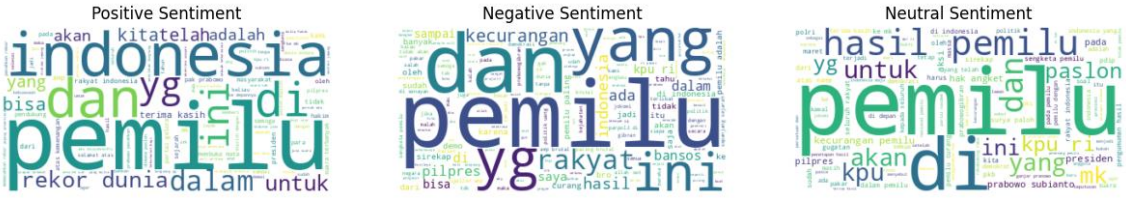

Figure 7. Frequent Words

4. CONCLUSION

The analysis and classification of public opinion on Twitter can be efficiently conducted using a pre-trained language model, specifically BERT, within the Google Colab environment. The successful application of the BERT model in analyzing the sentiment of 597 tweets within the dataset indicates that employing a pre-trained language model for sentiment analysis on Twitter is a promising approach in the realm of social media sentiment research. Despite its advantages, the data crawling process entails the necessity for hardware with substantial specifications. A notable limitation of this research is the constrained access to Twitter data, both in terms of temporal availability and the volume of data that can be extracted through the employed tools. The classified data will then be visualized with tools such as Matplotlib and Wordcloud to elucidate the data distribution.

5. SUGGESTED

This study continues to focus on BERT (Bidirectional Encoder Representations from Transformers) for language comprehension. However, there's potential for advancement by integrating the most effective pre-training approach for sign language recognition, incorporating tokenization methods as well (BEST) [10].

6. REFERENCES

[1] H. K. P. Hafiidh, Y. Andriani, D. A. Irawan, Supriatin, and N. T. Hidayat, "Twitter sentiment analysis classification to assess public opinion on football matches using the Naïve Bayes method," *Journal of Big Data*, vol. 2, no. 2, pp. 37-29, 2024.

[2] M. F. Fahrezi and A. A. Permana, "Sentimen analisis opini masyarakat pada sosial media Twitter terhadap organisasi aksi cepat tanggap menggunakan Naïve Bayes classifier," *Jurnal Teknik*, vol. 11, no. 2, pp. 113-121, 2022.

[3] D. A. Firdlous and R. Andrian, "Analisis sentimen publik Twitter terhadap Pemilu 2024 menggunakan model Long Short Term Memory," *SISTEMASI: Jurnal Sistem Informasi*, vol. 12, no. 1, pp. 52-60, 2023.

[4] J. A. Pratama, Y. Suprijadi, and Zulhanif, "Analisis sentimen sosial media Twitter dengan algoritma machine learning menggunakan software R," *Jurnal Fourier*, vol. 6, no. 2, pp. 85-89, 2017.

[5] S. Samsir, D. Irmayani, F. Edi, J. M. Harahap, Jupriaman, R. K. Rangkuti, B. Ulya, and R. Watrianthos, "Naïve Bayes algorithm for Twitter sentiment analysis," *Journal of Physics: Conference Series*, vol. 1933, no. 1, p. 012019, 2021. DOI: 10.1088/1742-6596/1933/1/012019.

[6] N. Siby and G. Joseph, "Twitter sentimental analysis using ML," in *Proceedings of the National Conference on Emerging Computer Applications (NCECA)*, vol. 3, no. 1, pp. 178-182, 2021, DOI: 10.5281/zenodo.5109050.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv* [cs.CL]. 2019.

[8] A. M. Mantika, A. Triayudi, and R. T. Aldisa, "Sentiment Analysis on Twitter Using Naïve Bayes and Logistic Regression for the 2024 Presidential Election", *SaNa: Journal of Blockchain, NFTs and Metaverse Technology*, vol. 2, no. 1, pp. 44–55, 2024.

[9] M. Pota, M. Ventura, R. Catelli, and M. Esposito, "An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian", *Sensors*, vol. 21, no. 1, 2021.

[10] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "BEST: BERT Pre-Training for Sign Language Recognition with Coupling Tokenization", *arXiv* [cs.CV]. 2023.