Implementation of Sentiment Analysis and Social Network Analysis of Twitter User Opinions on Web-Based Political Issues

Syair Muharram [1,] Asriyadi [2] and Muh. Ahyar [3]
[1,2,3] Departement of Electrical Engineering, Politeknik Negeri Ujung Pandang, Makassar, 90245, Indonesia
[*,b] asriyadi@poliupg.ac.id (Corresponding Author), [a] syairmuharram5@gmail.com, [c] ahyar@poliupg.ac.id

***Abstract*—The functionality of both sentiment analysis and social network analysis on social media such as Twitter or X can holds significant potential for understanding public opinion, especially in political-related issues. Despite its potential, there is currently lack of application that can provide these types of analyses. This research aimed to help to analyse public opinion in the field of political issues by using sentiment analysis and social network analysis techniques that can help to analyse the tendency of public opinion polarity on Twitter or X social media and to visualize the relationship of social media accounts with other social media accounts. The results obtained from this research are web applications that can help in problems related to sentiment analysis and network analysis using libraries from the Python programming language, and NodeJs. Based on the results of blackbox testing and technical testing, it can be concluded that the system that has been built can run in accordance with the predetermined design.**

Keywords—Sentiment Analysis; LSTM; Social Network Analysis; Twitter; Web

I. Introduction

The development of information technology today brings many advantages and benefits to society. It can be made more easier for people to interact with many people, expanding relationships, distance and time are no longer a problem, it is easier to express opinion themselves, and information can be spread quickly. Based on research conducted by Wearesocial Hootsuite in January 2019, social media users reach 150 million of the total population, whereas for the gadget users reach 130 million or about 48% of the total population[1].

Social media users can express their opinions on various issue topics, for example on political issues or other issues on social media platforms. One of the social media platforms that can be used as a medium for opinion is Twitter or X. Twitter is a messaging service that share similar feature with other communication tools. It has similar element with email, IM, SMS, blogging, RSS, etc[2].

On the developer side, Twitter offers its Application Programming Interface (API) services that allow other developers to obtain Twitter data for further research[3]. This making Twitter often use for data mining processes. This research uses two of data mining application, Sentiment Analysis and Social Network Analysis. Due to the updated policy published by Twitter by 9 February 2023 regarding the usage of APIs to be limit and paid[4], therefore the steps on this research use third-party tools for scraping the data.

Sentiment analysis is a part of opinion mining[5], it is a process to determine whether a sentence contains positive, negative or neutral tendencies[6]. Many methods can be used to form a sentiment analysis, one of them is lexicon-based method. Lexicon based method can be used to applying label into the data for using to training LSTM Model[7]. LSTM or Long Short-Term Memory is used to overcome the vanishing gradient problem[8]. LSTM can extract words and sentences with different contributions, and combine LSTM's embedding technology to classify text.

Beside sentiment analysis above, there is social network analysis that is often used to determine the central actor in a network by calculating the centrality value[9]. Actors are often known as nodes and the relationship between actors are known as edges. The relationship amongst actors can provide an insight on the social phenomenon within the networks[10]. Social network analysis's visual representation can be made by using graph analysis techniques[11]. There are various methods in building social networks based on centrality calculation, one of the calculation methods is degree centrality.

These two analyses can be combined to evaluate the public opinion especially in the field of politics including to mapping public sentiment about political situation, such as election in Indonesia with specific keyword and to know how it is connected to social interactions. But unfortunately, for now there is a lack of applications to help in the things mentioned above. The development of this website uses the waterfall model with performance testing conducted through blackbox testing.

II. Research Methodology

In this research, a website is created by referring to one of the SDLC models, which is the waterfall model. Waterfall models have 5 stages, the stages are: analysis or requirement, design, implementation, testing, and maintenance[12].

A. Waterfall Model

The waterfall model is suitable for system creation projects as well as large-scale system or software development[13].

1) Requirement: This phase serves as starting phase where the requirements essential for the work process are examined. Requirement in the data retrieval process are third-party scraping tool, which can save the data file in CSV format. Total of data collected are 2625 data. Furthermore, in this research, a lexicon-based approach is used for sentiment analysis. The lexicon used in this research is InSet Lexicon[14].

2) *Design:* In the design phase, some flowchart has been used in this research. Figures 1, 2, and 3 depict the stages of sentiment analysis, social network analysis, and website workflow.

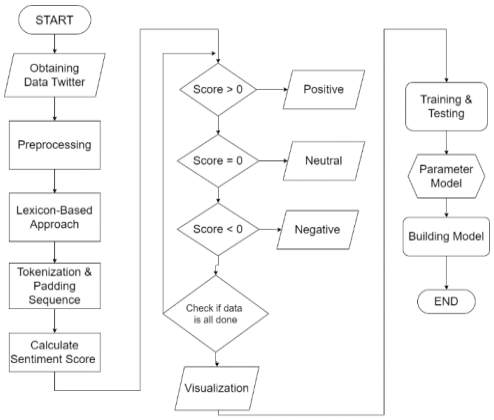

Figure 1. Sentiment Analysis Flowchart

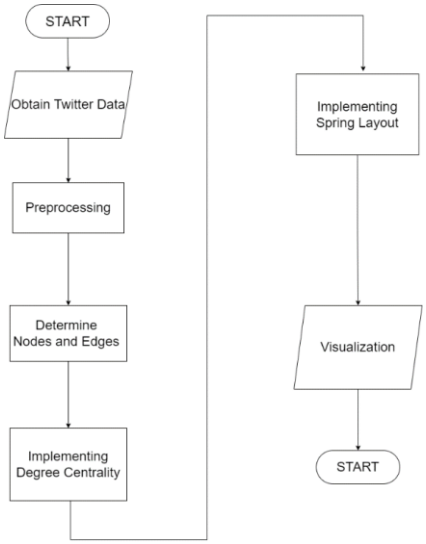

Figure 2. Network Analysis Flowchart

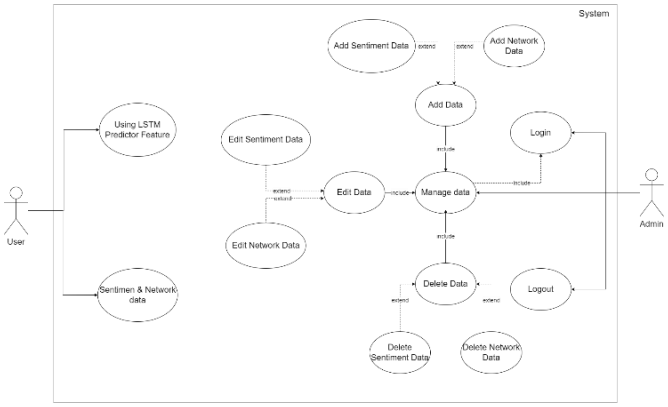

Figure 3. Use Case Diagram

3) *Implementation*: Several frameworks and programming languages are used in this research. Node.js and Flask are used for backend development, while Next.js is used for the frontend, and MySQL is used for the database. The LSTM model is created using TensorFlow, and NetworkX is used for the visual representation of social network analysis.

4) *Testing*: At this stage, Blackbox testing is utilized to test the program, with the expectation that the designed web application can operate as intended.

5) *Maintenance*: The goal of this maintenance stage is to optimize the software-related system that has been designed.

B. Blackbox Testing

Black box testing relies on the specialization of the software functions. The conducted testing can provide input conditions and perform tests on the specialization of the software's functions. The blackbox testing method

consists of several approaches; the blackbox method applied in this research is equivalence partitioning.

III. Results and Discussion

Result in this research divide into five parts, namely LSTM, sentiment analysis, social network analysis, design website, and testing. Each part is explained below:

A. LSTM

This research produces LSTM model that can be use to predict sentiment of tweet data. Tweet data are gathered with third-party tools with output of the tools is csv file format containing the following columns: username, handle, postdate, responding, retweets, and likes.

The preprocessing stage usage is to clean the raw tweet data that has been previously collected. First, data cleaning is performed to remove an unnecessary element, such as #, @, and the phrase "replying to" In addition to cleaning the data, it also converts sentences from uppercase to lowercase. The next step is tokenization, which breaks down sentences into tokens. Data labeling process is done by matching the words resulting from tokenization against the lexicon and summing the total score obtained from the sentiment values of each word in the lexicon. An example can be seen in Table 1 below.

Table 1. Applied Lexicon Labelling

Preprocessed	Tokenization	Score	Sentiment
pahamkan ya siapa yang bohong anies yang meminta bukan demokrat yang memaksa pd bongkar fakta	['paham', 'ya', 'bohong', 'anies', 'demokrat', 'paksa', 'pd', 'bongkar', 'fakta']	-15	negative
siapapun wakilnya yang penting presidenya anis baswedan ayo kita menangkan	['wakil', 'presidenya', 'anis', 'baswedan', 'ayo', 'menang']	3	positive
maju untuk Indonesia	['maju', 'indonesia']	0	neutral
kalo ga sama rocky gerung harusnya anis baswedan sama orang ini aja	['kalo', 'ga', 'rocky', 'gerung', 'anis', 'baswedan', 'orang', 'aja']	0	neutral
gus muhaimin pemimpin yang visioner amanah komunikatif islam moderat dan indonesia lebih baik	['gus', 'muhaimin', 'pimpin', 'visioner', 'amanah', 'komunikatif', 'islam', 'moderat', 'indonesia']		

After labeling the 2625 tweet data, the next step is creating a model using the LSTM architecture. LSTM has been built by dividing dataset with 2625 data to 80:20 ratio with parameter model: dropout = 0.2, embed_dim = 32, hidden unit = 16, optimizer = Adam, learning rate = 0.01, epoch = 50, batch_size = 128, dan validation_split = 0.1. As seen on Figure 4, from the results of the training and testing, it was found that the model went through overfitting. Therefore, the next step to address the overfitting is to use the K-Fold cross-validation method during the model training process with parameters n_splits=5, shuffle=True, and random_state=42.

Figure 4. Training Accuracy

Figure 5. K-Fold Cross Validation

The results from the 5-fold training showed an average accuracy of around 71%, which was obtained by summing the accuracy of each fold and dividing by the number of folds as seen on Figure 5.

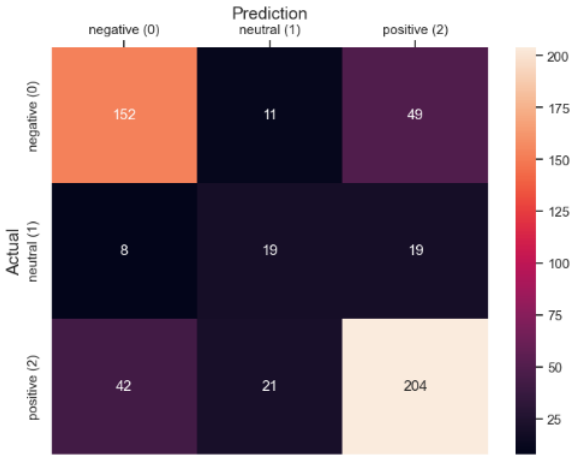

Figure 6. Confusion Matrix

For the confusion matrix as shown on Figure 6 above, with using 525 data test derived from 80:20 split. Accuracy is obtained for each class using the above formula is as follows: the negative class has an accuracy of 79%, the neutral class has an accuracy of 89%, and the positive class has an accuracy of 75%.

B. Sentiment Analysis

The preprocessing stage aims to eliminate unnecessary elements. Next, a pre-trained LSTM model will be used to perform sentiment prediction. The sentiment prediction results will then be used to create a sentiment percentage chart, a word cloud of the most frequently used words, a negative word cloud, and a positive word cloud.

C. Social Network Analysis

In the social network analysis, a preprocessing stage will also be conducted. The 'handle' column will be changed to the 'source' column, and the 'responding' column will be changed to the 'target' column to facilitate programming. These two columns will be used as nodes required for creating the social network analysis.

In the visualization stage of the social network analysis, the Python library NetworkX will be used. For the network analysis graph visualization, degree centrality with spring layout will be applied.

D. Design Website

Based on the previous stages of the waterfall model, the final result is a website design. There are five web pages, namely: Homepage, Login Page, Admin Dashboard, Sentiment Analysis Result Page, and Social Network Analysis Result Page.

1) Homepage

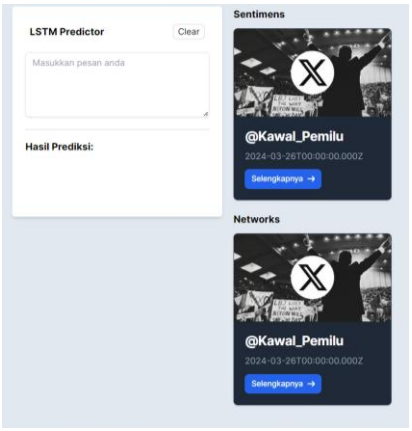

Figure 7. Homepage

The Homepage contains two cards: the Sentiments card, which has a 'More' button that navigates to the sentiment results page when clicked, and the Networks card, which has the same button that navigates to the network results page when clicked. Besides these two cards, this page also features an LSTM Predictor where users can input sentences they want to predict. This feature uses a pre-trained LSTM model.

2) Login Page

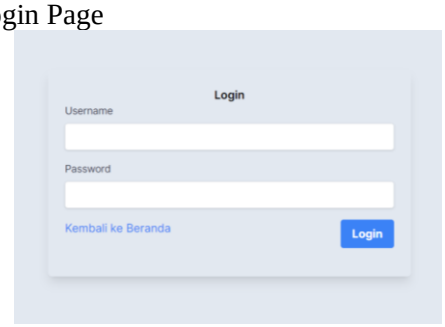

Figure 8. Login Page

The Login page can be accessed via the endpoint /login. After entering a username and password, the page will navigate to the admin dashboard. The admin dashboard allows users to upload raw data, which will then be automatically processed by the backend and at the same time send the data to the database when upload button is clicked. The dashboard page shown in Figure 9.

3) Admin Dashboard Page

Figure 9. Admin Dashboard

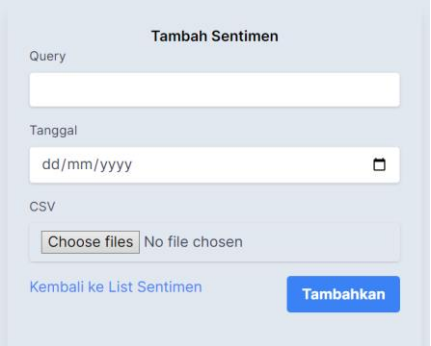

Figure 10. Upload Form

When adding the raw twitter data, admin must fulfill the query form, date, and csv file. Form in sentiment analysis and social network analysis has the same structure.

4) Sentiment Analysis Result Page

Figure 11. Sentiment Analysis Result Page

The Sentiment Analysis Result page has three visualizations: sentiment percentage, wordcloud of most frequently occuring word, and negative and positive word clouds. These results are obtained from the processing of uploaded tweet data.

5) Social Network Analysis Result Page

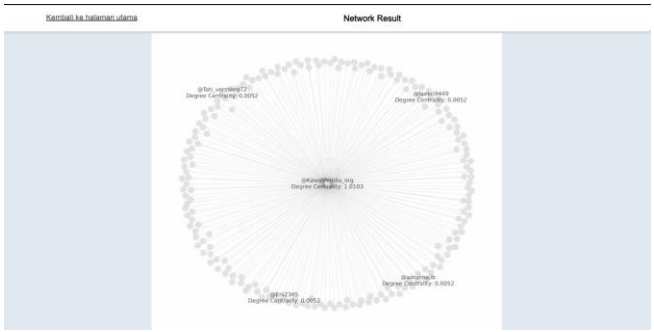

Figure 12. Network Analysis Result Page

The Social Network Analysis Result page presents a social network analysis visualization with degree centrality derived from the uploaded tweet data. The image shown in the social network analysis page is produced by NetworkX python library.

E. Webstite Testing

During the blackbox testing phase, we provided 7 test scenarios for the features on the created website, as shown in Table 2.

Table 2. Blackbox Testing

No	***Test Class***	***Expected Result***	***Description***
1	*Selengkapnya* button on sentiments & networks cards	Displays the Sentiment Analysis results page & Social Network Analysis results page.	Success
2	Accessing the login page	Displays the admin dashboard.	Success
3	Logout button	Returns to the homepage.	Success
4	Add data button on sentiment and networks pages	Adds data to the admin page, home page, and MySQL database.	Success
5	Edit button on sentiment and networks pages	Updates data on the admin page, home page, and MySQL database.	Success
6	Delete button on sentiment and networks pages	Deletes the uploaded data.	Success

No	*Test Class*	*Expected Result*	*Description*
7	Typing a sentence in the lstm predictor form	Displays the sentiment prediction result.	Success

Based on the blackbox testing conducted in this research, it can be concluded that the functionality of each feature works as expected. However, it should be noted that the sentiment analysis preprocessing feature has a relatively long response time of about 2-3 minutes for approximately 300-400 tweets, due to the data cleaning process.

IV. Conclusion

1. The model obtained during training experienced overfitting. Therefore, K-Fold cross-validation was performed, resulting in a training accuracy is 71%, and testing accuracy is 74%.
2. The Social Network Analysis is generated with degree centrality and spring layout by using NetworkX python library.
3. Based on blackbox testing used in this research showed that all features and functionalities on the website were working as intended.

Acknowledgement

My sincere gratitude to both of my parents and the Department of Electrical Engineering and Multimedia and Network Engineering Study Program for helping to arrange this research.

References

[1] I. Izzulsyah, A. Nur Hidayah, and L. Saputra, "ANALISIS PENGGUNAAN MEDIA SOSIAL DI MASA PANDEMI (ANALYSIS OF SOCIAL MEDIA USE DURING PANDEMIC)," Jurnal Fraction, vol. 2, no. 1, p. 21, 2022.

[2] M. G. Undap, V. P. Rantung, and P. T. D. Rompas, "Analisis Sentimen Situs Pembajak Artikel Penelitian Menggunakan Metode Lexicon-Based."

[3] S. M. Alzahrani, "Big Data analytics tools: Twitter API and spark," in 2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021, Institute of Electrical and Electronics Engineers Inc., Mar. 2021. doi: 10.1109/WIDSTAIF52235.2021.9430205.

[4] XDevelopers, "Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead," Twitter. Accessed: Apr. 21, 2023. [Online]. Available: https://x.com/XDevelopers/status/1621026986784337922

[5] F. T. Saputra, S. H. Wijaya, Y. Nurhadryani, and Defina, "Lexicon Addition Effect on Lexicon-Based of Indonesian Sentiment Analysis on Twitter," in Proceedings - 2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 136–141. doi: 10.1109/ICIMCIS51567.2020.9354269.

[6] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," J Big Data, vol. 5, no. 1, Dec. 2018, doi: 10.1186/s40537-018-0164-1.

[7] R. L. Mustofa and B. Prasetiyo, "Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter," in Journal of Physics: Conference Series, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1918/4/042155.

[8] S. Chakraborty, J. Banik, D. Chatterjee, and S. Addhya, "Study of Dependency on number of LSTM units for Character based Text Generation models," 2020.

[9] D. Inayah et al., "Seminar Nasional Official Statistics 2020: Statistics in The New Normal: A Challenge of Big Data and Official Statistics IMPLEMENTASI SOCIAL NETWORK ANALYSIS DALAM PENYEBARAN INFORMASI VIRUS CORONA (COVID-19) DI TWITTER (Implementation Social Network Analysis in Distribution of Corona Virus (Covid-19) Information on Twitter)."

[10] S. W. Wiiava and I. Handoko, "Examining a Covid-19 Twitter Hashtag Conversation in Indonesia: A Social Network Analysis Approach," in Proceedings of the 2021 15th International Conference on Ubiquitous Information Management and Communication, IMCOM 2021, Institute of Electrical and Electronics Engineers Inc., Jan. 2021. doi: 10.1109/IMCOM51814.2021.9377382

[11] L. C. Freeman, The development of social network analysis : a study in the sociology of science. Empirical Press, 2004.

[12] Y. Bassil, "A Simulation Model for the Waterfall Software Development Life Cycle," 2012. [Online]. Available: http://iet-journals.org/archive/2012/may_vol_2_no_5/255895133318216.pdf

[13] Aceng Abdul Wahid, "Analisis Metode Waterfall Untuk Pengembangan Sistem Informasi," Jurnal Ilmu-ilmu Informatika dan Manajemen STMIK, no. November, 2020.

[14] F. Koto and G. Y. Rahmaningtyas, "Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs," in 2017 International Conference on Asian Language Processing (IALP), 2017, pp. 391–394. doi: 10.1109/IALP.2017.8300625