ICEBIT
The 4th International Conference On Economics , Business and Information Technology
June 10, 2023 pp. 803-309

Sentiment Analysis of Twitter Discussions on Rafael Alun: Multinomial Naïve Bayes and Decision Tree Approach

Iqbal Sabilirrasyad
Institut Teknologi dan Sains Mandala
Jember, Indonesia
085171243269, +628
iqbal@itsm.ac.id

Zainul Hasan
Institut Teknologi dan Sains Mandala
Jember, Indonesia
082232425856, +628
zainulhasan@itsm.ac.id

Mas'ud Hermansyah
Institut Teknologi dan Sains Mandala
Jember, Indonesia
1330466463. +628
masudhermansyah@itsm.ac.id

ABSTRACT

Political events in Indonesia often draw significant attention on the social media platform Twitter, with netizens using the medium to express their opinions and feelings. One such viral topic on Twitter revolves around Rafael Alun, a former Director General of Taxation at the Indonesian Ministry of Finance who has been implicated in a possible gratification scandal and subsequently investigated by the Corruption Eradication Commission (KPK). As a result, Rafael Alun's name has been trending on Twitter. By using sentiment analysis, it is possible to identify the prevailing sentiment elements within tweets related to Rafael Alun. The application of Multinomial Naïve Bayes and Decision Tree algorithms serves to determine the accuracy of sentiment analysis results derived from Twitter tweets data. The research process involves steps such as data pre-processing, data processing, classification and evaluation. The sentiment analysis of tweets containing the mention of "Rafael Alun" shows that the majority of the tweets express a negative sentiment. The accuracy rates obtained by implementing the Multinomial Naïve Bayes and Decision Tree algorithms are 77% and 72% respectively. It is worth noting that these percentages are relatively moderate due to the unbalanced distribution of positive, negative and neutral sentiments on the topic.

Keywords : Sentiment Analysis, Multinomial Naïve Bayes, Decision Tree

1. INTRODUCTION

Rafael Alun was a former Director General of Taxation at the Indonesian Ministry of Finance who was the subject of an investigation by the Corruption Eradication Commission (KPK). Rafael Alun's son became the subject of widespread viral attention due to a violent incident several weeks earlier, sparking opinion about Rafael Alun's parenting skills and allegedly leading Rafael Alun to leave the Directorate General position. On 3 April 2023, the KPK arrested Rafael Alun Trisambodo, a former official at the Ministry of Finance's Directorate General of Taxes, for alleged gratification. This particular case has gained a considerable amount of traction on social media platforms, especially on Twitter, where individuals have been expressing their opinions and views on the matter. Twitter was always a platform for individuals to express their opinions on various cases such as politics, society, economy and culture have also become subjects of discussion and criticism among Indonesian netizens. The culture of Indonesian society has become ingrained with this phenomenon.

As one of the largest social media platforms, Twitter provides a platform for users to express their opinions and sentiments through features such as retweeting and directly replying to others' posts. Known for its microblogging and social networking capabilities, Twitter allows for in-depth tweets exploration of different topics and elements within messages, including hashtags. The accessibility of information through Twitter has contributed to its widespread popularity within the Indonesian netizen.

In the realm of mass media, Twitter offers insights into the diverse range of opinions on topics such as the Rafael Alun case and the allegations of gratification. but it's nevertheless difficult to determine what sentiment was given in every tweet on Twitter. With Sentiment Analysis, the overall sentiment around a trending topic on Twitter can be measured. This analytical approach has been used in numerous studies to assess the response to different trending topics on different platforms. For example, researcher Hermansyah M. explored public opinion on East Java POMPROV activities in 2022 through the social media platform Twitter. (Hermansyah, 2022) In addition, researcher Amilia explored public opinion on the anti-LGBT campaign through Twitter. (Fitri, Andreswari, & Hasibuan, 2019) It is worth noting that sentiment analysis is not limited to Twitter; other platforms such as YouTube have also been subjected to sentiment analysis, as seen in Novendri research about sentiment analysis on youtube about movie trailer (Novendri, Callista, Pratama, & Puspita, 2020). A variety of researchers have performed sentiment analysis using a variety of methods, including Naive Bayes, Decision Tree and Random Forest. (Fitri, Andreswari, & Hasibuan, 2019) (Hermansyah, 2022) The choice of method is not influenced by the specific topic and the range of opinions within it. Rather, it depends on how we want to approach the topic and what conclusions we want to draw from the data. We might get a different result on a same topic using the same method, because people have a different perspective on the given topic. For example, Villavicencio used Naïve Bayes to achieve an accuracy of 81.77% when analysing COVID- 19 sentiment twitter, while Wongkar achieved an accuracy of 80.90% using the same method to compare opinions between Indonesian presidential candidates Jokowi and Prabowo in 2020 (Villavicencio,

Macrohon, Inbaraj, Jeng, & Hsieh, 2021) (Wongkar & Angdresey, 2019) It is worth acknowledging that sentiment analysis has its own challenges given the variety of topics and discussions, since each topic offers a unique range of results that can be explored using the same method.

This study aims to perform a sentiment analysis on the case of Rafael Alun, focusing on several tweets that mention "Rafael Alun" as a keyword. By examining the sentiment expressed in these tweets, valuable insights can be gained. With thousands of posts on social media platforms every day, individuals have the freedom to openly express their opinions. These opinions include positive, negative and neutral sentiments towards a given topic. Positive sentiment indicates a favourable opinion towards rafael alun, negative sentiment conveys disapproval and negative opinion towards rafael alun, while neutral sentiment encompasses views that are impartial and outside the topic. Data pre-processing and data cleansing techniques are used to ensure accurate classification. In this research, sentiment analysis is performed using Naive Bayes, Decision Tree and Random Forest algorithms, each contributing to the analysis process.

2. Literature Overview

2.1 Sentiment analysis

Sentiment analysis is a hybrid approach that incorporates both data mining and text mining techniques. It serves as a valuable tool for extracting and analyzing diverse opinions expressed by consumers or experts across multiple media channels about products, services, or organizations. The primary objective of sentiment analysis is to comprehensively process textual data and extract the underlying sentiment embedded in an opinion. This sentiment can be classified into three main categories: positive, negative, and neutral. By using sentiment analysis, one can effectively examine and interpret the reactions and sentiments expressed by individuals through popular social media platforms.

Conducting surveys or using sentiment analysis can provide valuable insights into users' responses and reactions to the topic under discussion [Mas]. Such an approach allows researchers to gather information about users' feelings and opinions, thereby facilitating a deeper understanding of their perspectives. By analyzing the collected data, researchers can draw meaningful conclusions and make informed decisions based on the sentiments expressed by users. This method is proving to be an effective way to capture and analyze user feedback to gain valuable insights.

2.2 Twitter

Twitter serves as a prominent social media platform that facilitates real-time microblogging, allowing users to send short messages known as tweets (Wulandari, Saedudin, & Andreswari, 2021). The widespread use of Twitter transcends age barriers, with a significant portion of the population being proficient, active, and engaged on the platform (Sipayung, Maharani, & Zefanya, 2016). Each tweet is limited to 140 characters, which requires concise expression and often leads to the use of abbreviations or occasional misspellings. However, despite this limitation, Twitter serves as an invaluable resource for sharing ideas and disseminating brief, yet powerful information to a wide audience (Hasan Basri, 2017). The abundance of information available on Twitter makes it an ideal source for text classification.

2.3 Multinomial Naïve Bayes

Multinomial Naïve Bayes (MNB) is a probabilistic classification algorithm commonly used in natural language processing (NLP). It is an extension of the Naïve Bayes algorithm (1), which assumes that the presence of a particular feature in a class is independent of the presence of other features. In the case of Multinomial Naïve Bayes, it is specifically designed to handle discrete features, such as word counts or word frequencies, in text classification problems. It is widely used for tasks such as document classification, sentiment analysis, and spam filtering. (Susanti, Djatna, & Kusuma, 2017,),

$$Posterior\ Probability = \frac{Conditional\ Probability * Prior\ Probability}{Predictor\ Prior\ Probability} \quad (1)$$

$$P\left(\frac{A}{B}\right) = \left(\frac{P(A \cap B)}{P(B)}\right) = \frac{P(A) * P\left(\frac{B}{A}\right)}{P(B)}$$

The algorithm works by calculating the probability of a document belonging to a particular class given its features. In the context of NLP, the features typically correspond to the frequencies or counts of words in the document. The algorithm makes the assumption that the feature probabilities follow a multinomial distribution. To train the Multinomial Naïve Bayes classifier, the algorithm estimates the probabilities of each feature occurring in each class. This is done by counting the occurrences of each feature in the training data for each class and then applying smoothing techniques like Laplace smoothing to handle unseen features.

During the prediction phase, the algorithm calculates the posterior probability of each class given the observed features in a document using Bayes theorem. The class with the highest posterior probability is assigned as the predicted class for the document. Multinomial Naïve Bayes is a simple and computationally efficient algorithm, which makes it popular for text classification tasks. However, it assumes independence between features, which may not hold true in all cases. Despite this simplifying assumption, it often performs well in practice and serves as a good baseline model for many NLP applications. (Susanti, Djatna, & Kusuma, 2017,)

2.4 Decision Tree

Decision trees are constructed by recursively selecting the best feature and threshold to split the data based on impurity measures such as gini impurity(2) or information gain(3). The goal is to minimize impurity or maximize information gain. The prediction for a new instance is determined by traversing the tree based on its features and assigning the class label or predicted value associated with the leaf node reached (Suthaharan, 2016).

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2 \qquad (2)$$

$$Information\ Gain = E_{parent} - AvgE_{child} \qquad (3)$$

3. Methodology

The research involved topic formulation, literature overciew, data collection, preprocessing, splitting the data for Multinomial Naïve Bayes and Decision Tree modeling, and drawing conclusions from the sentiment analysis results focusing on sentiment opinion about Rafael Alun.

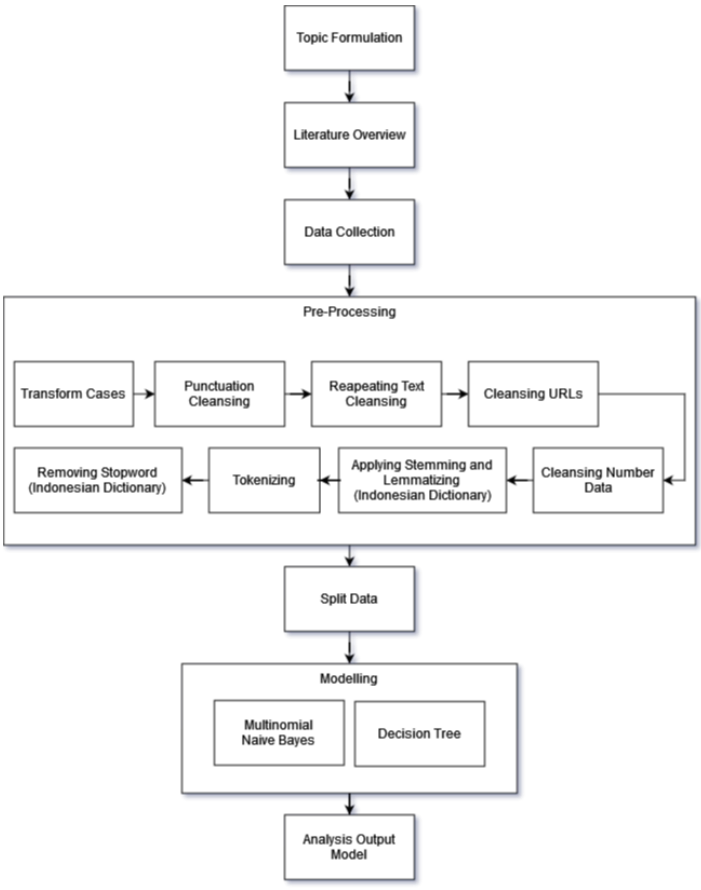

Figure 1 Methodology

The following is an explanation of the diagram table :

1. Topic Formulation : The research first involved the formulation of topics through an examination of trending topics related to Rafael Alun on Twitter.

2. Literature Overview : Relevant theories and previous research gathered to provide a solid foundation for the study.
3. Data Collecting : Process of collecting tweets data from Twitter using a Python-based web crawler that retrieves tweets containing the keyword "Rafael Alun" from search results starting from April 3, 2023, when Rafael Alun was arrested, until the research was conducted.
4. Pre-processing : Preprocessing plays a critical role in preparing raw data for machine learning algorithms. It includes a set of techniques and operations designed to address common challenges and improve the quality and suitability of the data for effective modeling. NLP data preprocessing includes steps such as tokenization, stop word removal, lowercase removal, punctuation removal, stemming/lemmatization, and URL handling. These steps cleanse and transform text data for NLP tasks..
    a. Case Transform: Change all existing tweet strings to lowercase. This is done so that existing text is consistent and can be more easily extracted and processed.
    b. Punctuation Cleansing : The action of removing punctuation marks from the text. Punctuation marks such as ',' (comma), '.' (period), ':' (colon), etc.
    c. Reapeating text Cleansing : the operation of eliminating the use of repeated words. Generally often used in the Indonesian language, such as Rudi Wijaya's tweet on April 28th

       “ Patut dipertanyakan juga dtkj yang bisa-bisanya memberikan usulan bodoh semacam ini. Apakah mereka pengguna transportasi publik atau jangan-jangan manusia hedon sejenis dengan rafael alun.”

       Repeated text cleansing ensures that only the first word is used if the word is repeated throughout the text.
    d. Cleansing URL’s : Some existing tweets typically have content in the form of website links or videos. During the data crawling process, these URLs are considered as plain text in the tweet data. The URL's were not providing any sentiments within the tweet. Therefore, cleansing URL's is needed even if the URL's filter feature from advance Search was applied.
    e. Cleansing Number Data: In the modeling process, we do not require numbers that come from a string of the tweet, the same as URL's to measure the level of sentiment on a topic, number does not provide any sentiment statement in the tweets.This process is performed to remove number data from the existing text.
    f. Applying Stemming and Lemmatizing (Dictionary Indonesia) : This process is performed to cut the existing text into the original root word structure. This process uses the Sastrawi module in Python to perform stemming and lemmatizing. The Sastrawi module known to provides a larger Indonesian vocabulary than other Python NLP modules. (Rosid, Fitrani, Astutik, Mulloh, & Gozali, 2020)
    g. Tokenizing : The tokenization process is a process of transforming existing text into smaller pieces called tokens. This process is performed after making each word in the text more standardized after the stemming and lemmatizing process.
    h. Removing Stopword (Dictionary) : The results of existing tokens are re-recognized to eliminate some words that are unnecessary or might even affect the modeling results. The use of stopword can reduce or increase the prediction results of a model. This is because the number of words that do not provide a sentiment value, which are commonly used in the Indonesian language, affect the prediction accuracy. This does not imply that the usage of stopwords will make the accuracy of a model become worse, but rather it can lead to better true-positive and false-positive accuracy (Pradana & Hayaty, 2019).
5. Split Data : The data is split into training and test sets before proceeding to the modeling stage. Different splits such as 6:4, 7:3, 8:2, and 9:1 are applied.
6. Modelling : Multinomial Naïve Bayes Classifier and Decision Tree algorithms are used for classification and analysis.
7. Analysis Output : The prediction results and accuracy are evaluated using the Confusion Matrix method, which provides insight into the classification performance by indicating the number of correct and incorrect predictions.

4. Result and Discussion

The research was conducted using Python programming language version 3.11. The entire research process involved several stages, starting with data collection through the use of web crawlers. The web crawlers were programmed to retrieve tweets containing the keyword "Rafael Alun" from Twitter, during the time between April 3 and May 28. As a result, a total of 1021 tweets were obtained for further analysis.

The collected tweets covered a wide range of topics related to Rafael Alun, including discussions about his case, corruption, and other related issues. The distribution of sentiments within the collected dataset was as follows: 433 tweets were classified as negative, indicating a critical or unfavorable sentiment; 134 tweets were classified as

positive, indicating a supportive or favorable sentiment; and 456 tweets were classified as neutral, indicating a lack of sentiment or an impartial viewpoint. Here is a graph of the distribution of target sentiment tweets.

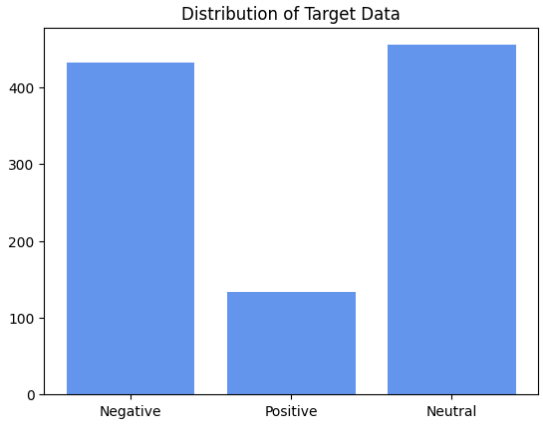

Figure 2 Distibution of Target Data

The data-target mapping process was performed manually by evaluating tweets based on their sentiment towards Rafael Alun. Negative targets were identified based on tweets containing negative elements or criticism towards Rafael Alun. Positive targets were determined by tweets expressing support or positive opinions about Rafael Alun. Neutral targets were assessed based on the overall topic discussed in the tweet. If the tweet focused on topics unrelated to Rafael Alun, then the intent of the Tweet was examined. By focusing only on tweets with clear positive or negative sentiment, the research aimed to improve the accuracy of sentiment classification specifically related to Rafael Alun. After the target mapping was completed, the data underwent a preprocessing stage.

To ensure the accuracy and relevance of the collected data, several modules and libraries were used, including Spacy and Sastrawi. These modules are specifically designed for text processing tasks in the Indonesian language, which was essential for this research. They provided advanced linguistic processing capabilities such as tokenization, lemmatization, and stopword removal to preprocess the raw text data.

Table 1. Progress preprocessing tweets result table

No	Preprocessing	Result
1	Tweet	Bagaimanapun DJP sudah sukses melahirkan 2 penjahat legendaris yang akan selalu dikenang masyarakat. Gayus Tambunan dan Rafael Alun Triambodo. . .
2	Transform Case	bagaimanapun djp sudah sukses melahirkan 2 penjahat legendaris yang akan selalu dikenang masyarakat. gayus tambunan dan rafael alun triambodo. . . .
3	Punctuation Cleansing	bagaimanapun djp sudah sukses melahirkan 2 penjahat legendaris yang akan selalu dikenang masyarakat gayus tambunan dan rafael alun triambodo
4	Repeating Text Cleansing	bagaimanapun djp sudah sukses melahirkan 2 penjahat legendaris yang akan selalu dikenang masyarakat gayus tambunan dan rafael alun triambodo
5	Cleansing URL's	bagaimanapun djp sudah sukses melahirkan 2 penjahat legendaris yang akan selalu dikenang masyarakat gayus tambunan dan rafael alun triambodo
6	Cleansing Number Data	bagaimanapun djp sudah sukses melahirkan penjahat legendaris yang akan selalu dikenang masyarakat gayus tambunan dan rafael alun triambodo
7	Applying Stemming and Lemmatizing	bagaimana djp sudah sukses lahir jahat legendaris yang akan selalu kenang masyarakat gayus tambun dan rafael alun triambodo
8	Tokenizing	bagaimana djp sudah sukses lahir jahat legendaris yang akan selalu kenang masyarakat gayus tambun dan rafael alun triambodo
9	Removing Stopword	bagaimana djp sukses lahir jahat legendaris akan selalu kenang masyarakat gayus tambun rafael alun triambodo

During the preprocessing stage, several techniques were applied to transform and to clean the data. The results of the preprocessing are presented in the following (table 1). The changes included converting all text to lowercase,

removing punctuation, addressing repetitive text, cleaning URLs from tweets, removing numeric data, applying stemming and lemmatization using the Sastrawi module, tokenizing the text into tokens, and removing stop words. These preprocessing steps resulted in a total of 11,140 word features extracted from the entire tweet dataset.

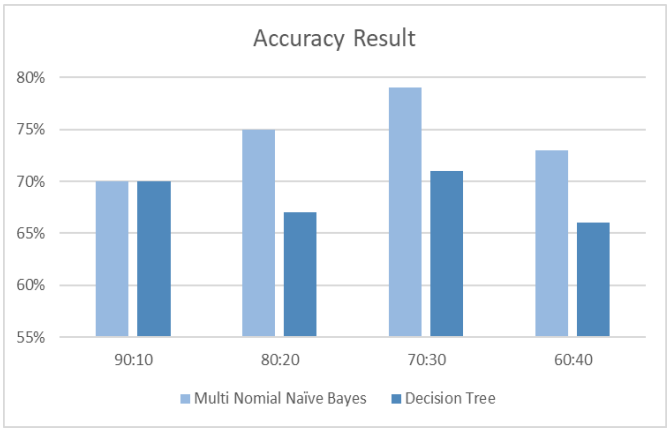

Figure 3 Accuracy Result Comparasion

The data splitting process was performed several times with different ratios of training data to test data: 40:60, 30:70, 20:80, and 10:90. This approach allowed for a comprehensive evaluation of the models' performance under different training and testing conditions. By examining the accuracy of the models across these different splits, we can gain insight into their efficiency in sentiment analysis on the topic of Rafael Alun. After splitting the data based on the predetermined ratios, the modeling and classification processes were conducted. The results showed interesting patterns in the accuracy of the models. In the 90:10 data split, where 10% of the data was being set for test, an overall accuracy of 70% was achieved. Specifically, the Multinomial Naïve Bayes model had a commendable accuracy rate of 75%, while the Decision Tree model had a slightly lower accuracy rate of 67%. Moving to the 80:20 data split, the Multinomial Naïve Bayes model showed improved performance with an accuracy of 79%. On the other hand, the Decision Tree model achieved an accuracy of 71%. For the 70:30 data split, where 30% of the data was allocated for testing, the Multinomial Naïve Bayes model showed a further improvement in accuracy, reaching 73%. Conversely, the Decision Tree model experienced a decrease in accuracy to 66%. Finally, in the 60:40 data split, the Multinomial Naïve Bayes model maintained a consistent accuracy of 73%, while the Decision Tree model's accuracy decreased slightly to 66%. These results indicate that the accuracy of the models varies depending on the ratio of training and testing data. In general, the Multinomial Naïve Bayes model showed superior performance compared to the Decision Tree model across the different data split scenarios, consistently achieving higher accuracy in sentiment classification for the given dataset.

5. Conclusion

The manual data-target mapping system allowed an additional focused evaluation of tweets with a clear positive or negative sentiment towards Rafael Alun. The next preprocessing step effectively transformed the raw text records, resulting in 11,140 word features. The splitting process allowed the Multinomial Naïve Bayes and Decision Tree models to be evaluated under specific training and testing conditions. The accuracy of the models varied depending on the splitting ratio, with the Multinomial Naïve Bayes version always outperforming the Decision Tree model at each different ratio. The results of the study suggest that sentiment analysis of Twitter data can provide valuable insights into public opinions and reactions to specific topics. The preprocessing techniques applied in the research contributed significantly to improving the quality and suitability of the data for modeling. In addition, the evaluation of different sentiment distribution ratios highlights the importance of data balancing and the use of some balancing data algorithms when evaluating the overall performance of sentiment analysis. Future research in this area can explore additional strategies and algorithms to further improve the accuracy and effectiveness of social media sentiment analysis. Further research can also focus on analyzing the patterns and trends in public sentiment closer to specific individuals or activities, along with other trending cases, to gain deeper insights into the dynamics of public opinion.

6. REFERENCES

Fitri, V. A., Andreswari, R., & Hasibuan, M. A. (2019). Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm. *Procedia Computer Science*(161), 765-772. doi:10.1016/j.procs.2019.11.181

Hasan Basri, S. (2017, Oktober). PERANMEDIA SOSIAL TWITTER DALAM INTERAKSI SOSIAL PELAJAR SEKOLAH MENENGAH PERTAMA DI KOTA PEKANBARU (studi kasus pelajar SMPN 1 kota Pekanbaru). *Jom FISIP, 4*(2), 1-15. Retrieved from https://jom.unri.ac.id

Hermansyah, M. (2022, November). ANALISIS SENTIMEN TWITTER UNTUK MENGETAHUI KESAN MASYARAKAT TENTANG PELAKSANAAN POMPROV JAWA TIMUR TAHUN 2022 DENGAN

PERBANDINGAN METODE NAÏVE BAYES CLASSIFIER DAN DECISION TREE BERBASIS SMOTE. *JURNAL JITEK, 2*(3), 249-255. doi: https://doi.org/10.55606/jitek.v2i3.551

Novendri, R., Callista, A. S., Pratama, D. N., & Puspita, C. E. (2020, June 26). Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes. *Bulletin of Computer Science and Electrical Engineering, 1*(1), 26-32. doi:10.25008/bcsee.v1i1.5

Pradana, A. W., & Hayaty, M. (2019, November). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 4*(4), 375-380. doi:http://dx.doi.org/10.22219/kinetik.v4i4.912

Rosid, M. A., Fitrani, A. S., Astutik, I. R., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, 1-6. doi:10.1088/1757-899X/874/1/012017

Sipayung, E. M., Maharani, H., & Zefanya, I. (2016, April). Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier. *Jurnal Sistem Informasi (JSI), 8*(1), 958-965. doi: https://doi.org/10.36706/jsi.v8i1.3250

Susanti, A. R., Djatna, T., & Kusuma, W. A. (2017, September). Twitter's Sentiment Analysis on Gsm Services using Multinomial Naïve Bayes. *Telkomnika (Telecommunication Computing Electronics and Control), 15*(3), 1354-1361. doi:10.12928/TELKOMNIKA.v15i3.4284

Suthaharan, S. (2016). Decision Tree Learning. *Integrated Series in Information Systems, 36*, 237–269. doi:https://doi.org/10.1007/978-1-4899-7641-3_10

Villavicencio, C., Macrohon, J., Inbaraj, X., Jeng, J., & Hsieh, J. (2021). Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes. (A. Zubiaga, Ed.) *Information*(12), 204. doi:https://doi.org/10.3390/info12050204

Wongkar, M., & Angdresey, A. (2019, October 16). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. *International Conference on Informatics and Computing (ICIC)*, 1 - 5. doi:10.1109/ICIC47613.2019.8985884

Wulandari, D. A., Saedudin, R. R., & Andreswari, R. (2021 , Oktober). Analisis Sentimen Media Sosial Twitter Terhadap Reaksi Masyarakat Pada Ruu Cipta Kerja Menggunakan Metode Klasifikasi Algoritma Naive Bayes. *e-Proceeding of Engineering, 8*(5), 9007. Retrieved from https://openlibrarypublications.telkomuniversity.ac.id