Sentiment Analysis of Cak Nun on Youtube and Online News: Multinomial Naive Bayes for Positive, Neutral, and Negative Perspectives

Moh. Heri Susanto[1]
[1] Universitas Islam Negeri Maulana Malik Ibrahim, Jalan Gajayana 50, malang, Indonesia
heribangkal21@gmail.com

Abstract

In the era of digitalization, which is saturated with abundant information, sentiment analysis has emerged as a crucial tool for comprehending and addressing the intricate nature of public opinion. The copious quantity of textual data generated by media platforms such as YouTube and online news outlets offers valuable insights into the viewpoints held by the public on a diverse range of topics and public figures. Consequently, sentiment analysis plays a pivotal role in discerning the prevailing direction of sentiment, be it positive, negative, or neutral. This article focuses on the sentiment analysis of Cak Nun, a highly esteemed cultural figure and poet from Indonesia. The utilized data consists of titles from YouTube and pertinent articles from online news sources that pertain to Cak Nun. The chosen methodology is the Multinomial Naive Bayes with CountVectorizer feature selection. By employing the Multinomial Naive Bayes, the patterns present within the text are learned to classify the textual data. At the same time, the CountVectorizer identifies the critical aspects within the evaluations of YouTube titles and online news articles. The resulting accuracy achieved is 82.11%, thereby indicating the effectiveness of the Multinomial Naive Bayes in accurately classifying sentiment. Overall, the model produces favorable outcomes, although there remains room for improvement in its ability to handle negative sentiment, as the precision, recall, and F1-Score values for negative sentiment are slightly lower than those for other sentiments. This sentiment analysis is anticipated to yield considerable advantages for the public figure known as 'Cak Nun,' as well as for the wider public and researchers in terms of further advancements and progress in the future.

Keywords: Sentiment analysis, Cak Nun, Naive Bayes, CountVectorize, Youtube, Online News

INTRODUCTION

In the contemporary era of abundant information in the digital realm, platforms dedicated to disseminating information, such as YouTube and News Online, play a pivotal role in shaping the opinions and perspectives of individuals. The proliferation of content on YouTube and online news platforms can readily influence numerous individuals' perceptions and viewpoints. Conversely, a considerable segment of the population still struggles to discern whether content is contentious or meritorious. Consequently, the need arises for a sentiment analysis system that can offer an objective assessment, enabling the public to comprehend thoroughly by precluding the harmful effects of content or news items that are negative, such as spurious news, critiques of ideas or ideologies, personal slights, repudiation of political viewpoints, and suspicions, among others.

One notable personality frequently garners attention is Cak Nun, a renowned Indonesian poet and intellectual. Cak Nun has acquired recognition for his inspirational creations and profound philosophical perspectives on life. While many individuals hold his contributions in high regard, akin to the media realm, public opinions of these public figures can diverge significantly.

YouTube, one of the foremost platforms for video sharing and social interaction [1], facilitates the influence of content on its platform without necessitating actual verification. Within this context, this article aims to elucidate the potential of sentiment analysis in comprehending the positive, neutral, and negative perceptions of Cak Nun on YouTube. Cak Nun transcended the realm of art and culture, frequently engaging with social and political issues. Consequently, the public's perception of him encompasses an appreciation for his artistic endeavors and an evaluation of his ideological stances.

Furthermore, this article will expound upon how the outcomes of sentiment analysis can be applied within the news domain. News media is pivotal in disseminating information about public figures and current events. By comprehending the public's perception of these figures, news media can enhance the caliber and relevance of their news coverage, thereby furnishing their readers with more informative and accurate reports. Consequently, sentiment analysis is emerging as an indispensable tool in the ever-evolving landscape of media in the digital era.

METHODS

RESEARCH DESIGN

Several stages that necessitate attention are depicted in Figure 1.

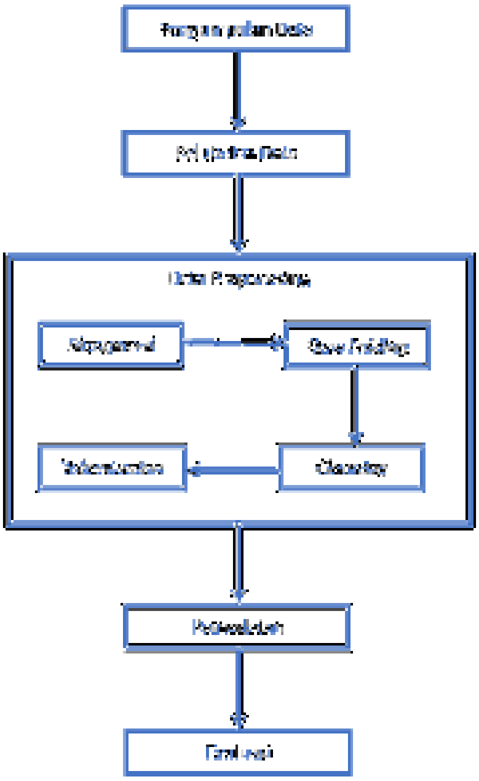


Figure 1. Research Design

DATA COLLECTION

Data collection is conducted during the Field Work Practice (PKL) at PT. Arion Indonesia. The data encompasses 1226 samples, with 231 being positively labeled, 62 being negatively labeled, and 933 being neutrally labeled. This comprehensive dataset is gathered from various sources such as YouTube and online news platforms. The graphical representation in Figure 1 illustrates the proportion of the data set.

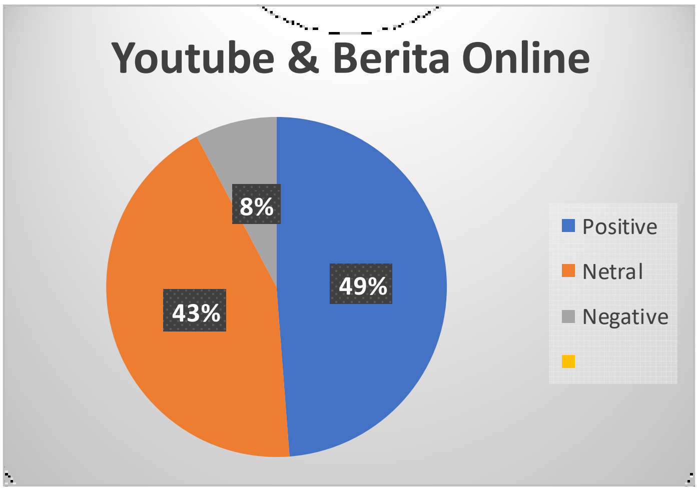


Figure 2. The Proportion of Data Collection

Data utilization before labeling can be seen in Table 1.

Table 1. Data collection results

No	SOURCE TYPE	TITLE	Manual Sentiment
1	Youtube	MALUNYA SAMPEK KE UBUN^2❗JOKOWI "FIR'AUN" JENGUK CAK NUN, HANYA JOKOWI YANG BISA❗	Negative
2	Youtube	"Materialisme: Perspektif Cak Nun dalam Melihat Ideologi Kontemporer"	Netral
3	Berita Online	Prabowo Doakan Cak Nun yang Mengalami Pendarahan Otak: Semoga Cepat Sembuh	Positive

SENTIMENT CATEGORIZATION

The categorization of sentiment holds significant significance in the domain of sentiment analysis. Considering the contextual factors, this study assigned a sentiment label to each YouTube content title and Online News. Researchers or research teams conduct the attribution of these labels with a specific focus on discerning and expressing the sentiment contained within each title. This labeling procedure was manually implemented to guarantee objectivity in sentiment determination. Figure 3 is a depiction of the process of categorizing sentiment.

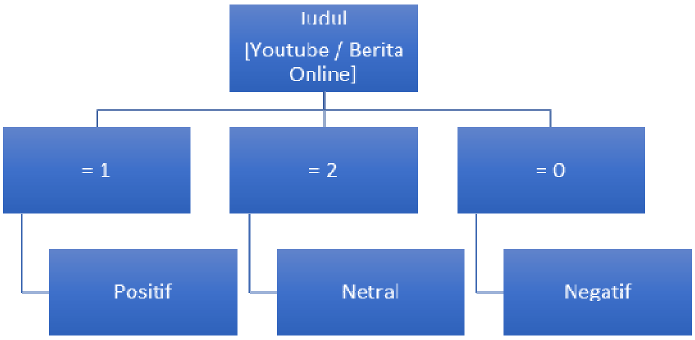


Figure 3. Sentiment Categorization Process

After the data collection process, the labeling data is illustrated in Table 2.

Table 2. Data Labeling Results

SOURCE TYPE	TITLE	Manual Sentiment	Label Sentiment
Youtube	MALUNYA SAMPEK KE UBUN^2❗JOKOWI "FIR'AUN" JENGUK CAK NUN, HANYA JOKOWI YANG BISA❗	Negative	0
Youtube	"Materialisme: Perspektif Cak Nun dalam Melihat Ideologi Kontemporer"	Netral	2
Berita Online	Prabowo Doakan Cak Nun yang Mengalami Pendarahan Otak: Semoga Cepat Sembuh	Positive	1

Only two characteristics were required in this investigation, specifically designated as 'TITLE' and 'Sentiment Label.'

PRE-PROCESSING DATA

The subsequent stage involves the pre-processing of the data. Consequently, data cleansing is executed to simplify the data for progression to the subsequent phase. Typically, the data that undergoes the pre-processing stage will differ from the data that does not. The data that undergoes pre-processing is superior as it can reduce memory usage, attain a structured format, and accelerate performance optimization.

The pre-processing stages encompass slang word transformation, case folding, cleansing, and tokenization. Slang word transformation endeavors to substitute a non-standard word with the closest word from the dictionary [2]. Case folding endeavors to convert all letters into lowercase [3]. Cleansing aims to eradicate characters or symbols such as link URLs (http://website.com), usernames or mentions (@username), hashtags (#), retweets, and emoticons [4]. Tokenization strives to segment a sentence by isolating each word present [5].

Table 3. Pre-processing Results

TITLE	Label Sentiment
[malunya, sampek, ke, ubun, jokowi, firaun, jenguk, cak, nun, hanya, jokowi, yang, bisa]	0
[materialisme, perspektif, cak, nun, dalam, melihat, ideologi, kontemporer]	2
[prabowo, doakan, cak, nun, yang, mengalami, pendarahan, otak, semoga, cepat, sembuh]	1

FEATURE EXTRACTION WITH COUNT VECTORIZER

Feature extraction is an advantageous technique for discerning diverse characteristics present in a document. In this instance, the countVectorizer, a tool provided by the scikit-learn library, transforms the text within X_train and X_test into a matrix that portrays the occurrence rate of words within said texts [6]. The process of count vectorizer computation is illustrated in Table 4 and Table 5.

Table 4. Building a Dictionary

3 TITLE Disatukan
[malunya, sampek, ke, ubun, jokowi, firaun, jenguk, cak, nun, hanya, yang, bisa, materialisme, perspektif, dalam, melihat, ideologi, kontemporer, prabowo, doakan, mengalami, pendarahan, otak, semoga, cepat, sembuh]

Table 5. Counting words and forming vectors

TITLE	Matriks
[malunya, sampek, ke, ubun, jokowi, firaun, jenguk, cak, nun, hanya, jokowi, yang, bisa]	[1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[materialisme, perspektif, cak, nun, dalam, melihat, ideologi, kontemporer]	[0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]
[prabowo, doakan, cak, nun, yang, mengalami, pendarahan, otak, semoga, cepat, sembuh]	[0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1]

NAÏVE BAYES CLASSIFICATION

The classification approach employed in this investigation was Naive Bayes. Naive Bayes is an algorithm that leverages Bayes' theorem and assumes that all attributes (words) in the dataset are independent [7]. Based on the training data, the Naive Bayes model examines the likelihood of word occurrences in each sentiment category (positive, neutral, negative). Additionally, this model is utilized to categorize sentiments from YouTube headlines and news stories into the corresponding sentiment categories. Equation (1) is the Naive Bayes formula [8], which can be computed for classification purposes.

$$P(A|B) = P(B|A)\ .\ P(A).\ P(B) \quad \ldots\ldots(1)$$

P (A|B) represents the Probability of A taking place given the occurrence of B, whereas P (B|A) signifies the probability of B happening given the proof that A has occurred. Similarly, P (A) denotes the Chance of A occurring, while P (B) corresponds to the Chance of B occurring.

The following are the sequential phases involved in the classification process of naive Bayes.

Calculating Priors

$$P(\text{negative}) = \frac{\textit{number of negative document}}{\textit{total document}} = \frac{1}{3} = 0{,}333$$

$$P(\text{positive}) = \frac{\textit{number of positive document}}{\textit{total document}} = \frac{1}{3} = 0{,}333$$

$$P(\text{neutral}) = \frac{\textit{number of neutral document}}{\textit{total document}} = \frac{1}{3} = 0,333$$

Calculating Likelihood

$$P(\text{word}|\text{class}) = \frac{\text{number of occurrences of "word" in the document is negative} + 1}{\text{total number of "words" in the negative document} + \text{number of unique words in the dictionary}}$$

$P(\text{malunya}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39}$

$P(\text{sampek}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39}$

$P(\text{ke}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39}$

$P(\text{ubun}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39}$

$P(\text{jokowi}|\text{negative}) = \frac{2+1}{13+26} = \frac{3}{39}$

$P(\text{ke}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39}$

...

$P(\text{semoga}|\text{negative}) = \frac{0+1}{13+26} = \frac{1}{39} = 0,025$

$P(\text{cak}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39} = 0,051$

$P(\text{nun}|\text{negative}) = \frac{1+1}{13+26} = \frac{2}{39} = 0,051$

$P(\text{selalu}|\text{negative}) = \frac{0+1}{13+26} = \frac{1}{39} = 0,025$

$P(\text{sehat}|\text{negative}) = \frac{0+1}{13+26} = \frac{1}{39} = 0,025$

$P(\text{dan}|\text{negative}) = \frac{0+1}{13+26} = \frac{1}{39} = 0,025$

$P(\text{bahagia}|\text{negative}) = \frac{0+1}{13+26} = \frac{1}{39} = 0,025$

...

$P(\text{semoga}|\text{positive}) = \frac{1+1}{11+26} = \frac{2}{37} = 0,054$

$P(\text{cak}|\text{positive}) = \frac{1+1}{11+26} = \frac{2}{37} = 0,054$

$P(\text{nun}|\text{positive}) = \frac{1+1}{11+26} = \frac{2}{37} = 0,054$

$P(\text{selalu}|\text{positive}) = \frac{0+1}{11+26} = \frac{1}{37} = 0,027$

$P(\text{sehat}|\text{positive}) = \frac{0+1}{11+26} = \frac{1}{37} = 0,027$

$P(\text{dan}|\text{positive}) = \frac{0+1}{11+26} = \frac{1}{37} = 0,027$

$P(\text{bahagia}|\text{positive}) = \frac{0+1}{11+26} = \frac{1}{37} = 0,027$

...

$P(\text{semoga}|\text{neutral}) = \frac{0+1}{8+26} = \frac{1}{34} = 0,029$

$P(\text{cak}|\text{neutral}) = \frac{1+1}{8+26} = \frac{2}{34} = 0,058$

P(nun|neutral) = $\frac{1+1}{8+26}$ = $\frac{2}{34}$ = 0,058

P(selalu|neutral) = $\frac{0+1}{8+26}$ = $\frac{1}{34}$ = 0,029

P(sehat|neutral) = $\frac{0+1}{8+26}$ = $\frac{1}{34}$ = 0,029

P(dan|neutral) = $\frac{0+1}{8+26}$ = $\frac{1}{34}$ = 0,029

P(bahagia|neutral) = $\frac{0+1}{8+26}$ = $\frac{1}{34}$ = 0,029

Menghitung Posterior

Data Testing: Semoga Cak Nun selalu sehat dan bahagia.

P(negative)

= negative Prior . P(semoga|negatif) . P(cak|negatif). P(Nun|Negatif) . P(selalu|Negatif) . P(sehat|negatif) . P(dan|negatif) . P(bahagia|Negatif)

= 0,333 . 0,025 . 0,051 . 0,051 . 0,025 . 0,025 . 0,025 . 0,025

= 0.000000000000845833008

P(positive)

= Positive Prior . P(semoga|positif) . P(cak|positif). P(Nun|positif) . P(selalu|positif) . P(sehat|positif) . P(dan|positif) . P(bahagia|positif)

= 0,333 . 0,054 . 0,054 . 0,054 . 0,027 . 0,027 . 0,027 . 0,027

= 0.00000000003

P(neutral)

= neutral Prior . P(semoga|netral) . P(cak|netral). P(Nun|netral) . P(selalu|netral) . P(sehat|netral) . P(dan|netral) . P(bahagia|netral)

= 0,333 . 0,029 . 0,058 . 0,058 . 0,029 . 0,029 . 0,029 . 0,029

= 0.00000000002

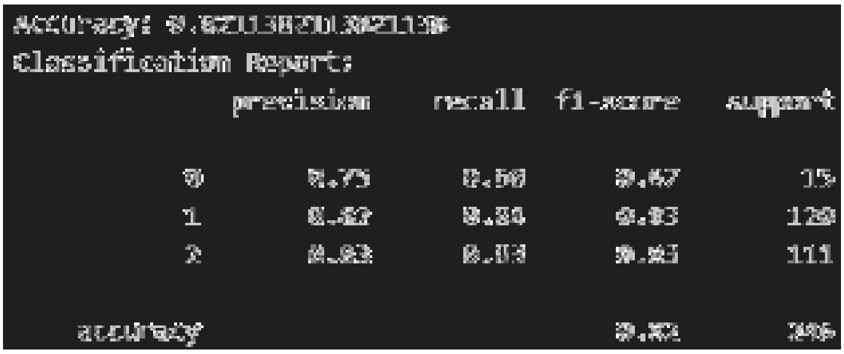
After obtaining the posterior calculation result in each class, the classification result of Naive Bayes is determined by selecting the highest value from these results. Among the calculated values, 0.00000000003 is found to be the most considerable value. Hence, this value can be utilized as the outcome for the classification of Naive Bayes. Consequently, when applying this calculation to the sentence *"Semoga Cak Nun selalu sehat dan bahagia."* the resulting classification is positive.

EVALUATION

An evaluation stage was conducted to ensure the integrity and effectiveness of the Naive Bayes classification model. This process entails utilizing evaluation metrics: accuracy, precision, recall, and F1-score. The results derived from this assessment aid in ascertaining the competency of the Naive Bayes algorithm in categorizing emotions conveyed in comments and news articles posted on the video-sharing platform YouTube. Through a thorough evaluation, it becomes feasible to gauge the extent to which the model can accurately portray positive, neutral, and negative perspectives regarding Cak Nun within the analyzed dataset.

RESULTS AND DISCUSSION

In sentiment analysis, a confluence occurs between YouTube content titles and Online News about Cak Nun. The output illustrated in Figure 4 is derived after undergoing the sequential data collection, pre-processing, and feature extraction processes utilizing a count vectorizer, Naive Bayes classification, and validation.



```
Accuracy: 0.8211382113821138
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.60      0.67        15
           1       0.82      0.84      0.83       120
           2       0.83      0.83      0.83       111

    accuracy                           0.82       246
```

Figure 4. Model Accuracy and Evaluation Results

The precision outcome from the sentiment analysis conducted on the titles of YouTube content and news titles pertaining to Cak Nun demonstrates a numerical value of 0.8211382113821138, approximately 82.1%. This finding illustrates that the Naive Bayes model employed in this investigation successfully discerns sentiments with a commendable level of precision. Within this delineation analysis:

- Precision is an evaluation metric that gauges the accuracy of a model's positive predictions. This assessment is derived from a specific calculation.

  $$Precision = \frac{TP}{TP+FP}$$

  True Positive (TP): The quantity of accurate positive identification outcomes.

  False Positive (FP): The quantity of erroneous negative identification outcomes.

  High precision signifies that the judgment tends to be accurate when the model asserts that a title pertains to a specific category (positive, neutral, or negative).

- Recall quantifies the degree to which a model can detect and assign all occurrences of a particular classification. The computation is executed as follows:

  $$Recall = \frac{TP}{TP+FN}$$

  False Negative (FN) refers to the count of positive instances that were supposed to be detected but were not.

  A high recall rate demonstrates that the model did not overlook many instances that should have been included in that classification.

- The F1 score represents a unified metric that encompasses both precision and recall. The following is the formula for computing this score:

  $$F1\text{-}Score = 2\ x\ \frac{PrecisionxRecall}{Precision+Recall}$$

  A high F1 score signifies that the model can attain an equilibrium between precision and recall.

The findings from the classification report revealed that the model exhibited a commendable level of precision and recall, particularly in its ability to accurately classify titles into the "neutral" and "positive" categories. Although the "negative" category displayed a slightly lower recall rate, overall, the model yielded satisfactory outcomes in examining sentiment within YouTube content titles and news titles of Cak Nun.

CONCLUSION

The findings of a sentiment analysis that integrates YouTube and online news content titles have indicated that individuals generally hold positive views towards Cak Nun. A considerable proportion of titles from YouTube and online news content express a positive perspective on Cak Nun, which suggests that he exerts a positive influence across various media platforms.

Nevertheless, it is crucial to acknowledge the presence of neutral and negative viewpoints within these titles. The presence of neutral perspectives can be attributed to titles that are more descriptive or focused on conveying information. On the other hand, negative perspectives may stem from criticism or controversies surrounding Cak Nun in certain media content.

Moreover, the high level of accuracy achieved in this sentiment analysis underscores the reliability of the Naive Bayes model employed in sentiment classification. Despite slight variations in sentiment outcomes, this model yields satisfactory results in comprehending society's response to Cak Nun in different media contexts.

Notably, changes in public sentiment towards public figures like Cak Nun can be influenced by numerous factors, including the context of news or YouTube content, the topics discussed, and shifts in public attitudes. Consequently, it is imperative to interpret the results of this sentiment analysis within a broader framework, as they only provide a snapshot of the current scenario.

ACKNOWLEDGMENT

On this occasion, I express my profound gratitude to PT, specifically Arion Indonesia, and more specifically to Mr. Erik, Mr. Rinto, and Ms. Hida for their exceptional support and guidance throughout the research endeavor. The remarkable experience provided by the entire Arion team has significantly enhanced my knowledge and skills. It is essential to acknowledge the invaluable contribution of Dr. Totok Chamidy, a meticulous and instructive lecturer. Their contributions have played a crucial role in my professional development, and I sincerely appreciate this invaluable experience.

REFERENCES

[1] K. A. B. Permana, M. Sudarma, & W. G. Ariastina, “Analisis Rating Sentimen pada Video di Media Sosial Youtube Menggunakan STRUCT-SVM”, *Maj. Ilm. Teknol.* Elektro, 18(1), 113, 2019, DOI: https://doi.org/10.24843/MITE.2019.v18i01.P17.

[2] D. Alita, Y. Fernando, & H. Sulistiani, "Implementasi Algoritma Multiclass Svm Pada Opini Publik Berbahasa Indonesia Di Twitter", *jurnal tekno kompak*, vol. 14, no. 2, p. 86, 2020, https://doi.org/10.33365/jtk.v14i2.792.

[3] D. Darwis, N. Siskawati, & Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional", *jurnal tekno kompak*, vol. 15, no. 1, p. 131, 2021, https://doi.org/10.33365/jtk.v15i1.744.

[4] S. Styawati, N. Hendrastuty, & A. R. Isnain, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine", jurnal informatika jurnal pengembangan it, vol. 6, no. 3, p. 150-155, 2021, https://doi.org/10.30591/jpit.v6i3.2870.

[5] C. Misrun, E. Haerani, M. Fikry, & E. Budianita, "Analisis Sentimen Komentar Youtube Terhadap Anies Baswedan Sebagai Bakal Calon Presiden 2024 Menggunakan Metode Naive Bayes Classifier", *jurnal coscitech (computer science and information technology)*, vol. 4, no. 1, p. 207-215, 2023, https://doi.org/10.37859/coscitech.v4i1.4790

[6] R. Awangga and N. Khonsa', "Analisis Performa Algoritma Random Forest Dan Naive Bayes Multinomial Pada Dataset Ulasan Obat Dan Ulasan Film", *jurnal telekomunikasi dan komputer*, vol. 12, no. 1, p. 60, 2022, https://doi.org/10.22441/incomtech.v12i1.14770.

[7] C. Misrun, E. Haerani, M. Fikry, & E. Budianita, "Analisis Sentimen Komentar Youtube Terhadap Anies Baswedan Sebagai Bakal Calon Presiden 2024 Menggunakan Metode Naive Bayes Classifier", *jurnal coscitech (computer science and information technology)*, vol. 4, no. 1, p. 207-215, 2023, https://doi.org/10.37859/coscitech.v4i1.4790.

[8] A. A. Ningtyas, A. Solichin, & R. Pradana, "Analisis Sentimen Komentar Youtube Tentang Prediksi Resesi Ekonomi Tahun 2023 Menggunakan Algoritme Naïve Bayes", *bit (fakultas teknologi informasi universitas budi luhur)*, vol. 20, no. 1, p. 9, 2023, https://doi.org/10.36080/bit.v20i1.2317.