YouTube Comment Sentiment Analysis on Deddy Corbuzier and BEM UI's Podcast Using TF-IDF and Naïve Bayes

Muhammad Rafi Al Basyary

Data Science Study Program, Faculty of Science and Technology, Yogyakarta Technology University

Correspondence Author: ralbasyaroffice@gmail.com

Article Info :

Article History :

Received :
11 January 2024
Revised :
04 February 2024
Accepted :
03 July 2024
Available Online :
28 August 2024

Keyword :
Public Sentiment, Naïve Bayes Classifier, Youtube, Presidential Candidate Ideas, Sentiment Analysis

ABSTRACT

The development of technology and the popularity of digital platforms such as Youtube have had a significant impact on the dissemination of information, especially related to the idea of a presidential candidate in the 2024 election. For example, Deddy Corbuzier actively uploads a podcast video on his Youtube channel, together with the chairman of BEM UI, to discuss the ideas of each presidential candidate. In this study, sentiment analysis was carried out on more than 2.6 million Youtube user comments on the video using the Naïve Bayes Classifier algorithm. This algorithm has proven effective in previous studies, showing high accuracy in classifying people's sentiments. The research methodology includes data labeling, text preprocessing, word weighting with TF-IDF, data validation using k-fold cross validation, and data testing. The results of the sentiment analysis revealed that more than fifty percent of the comments were positive, while some remained neutral. The data visualization process using word cloud provides a clear picture of the topics most talked about by the public, with the word "Leader" dominating. Although the accuracy of the Naïve Bayes Classifier in this study reached 57.6%, this study provides valuable insights into the public's view of the idea of a presidential candidate. Further development may involve the use of other algorithms as comparators to improve the accuracy of sentiment analysis.

1. INTRODUCTION

Technological developments bring many benefits, one of which is the ease of accessing information via the internet. This is accompanied by the increasing growth of digital platforms such as Facebook, Instagram, Twitter, YouTube, and many more. The presence of YouTube is a new innovation because the information is presented in audio-visual form (Zhafira et al., 2021). The influence of the fast rate of information dissemination and the popularity of YouTube is currently being used by many parties as a platform for branding and publication, one of the content creators who is active in utilizing YouTube is Deddy Corbuzier. Deddy Coruzier uploaded a video podcast on his YouTube channel together with the chairman of BEM UI discussing the ideas of each presidential candidate in the 2024 elections.

KPU Chairman Hasyim Asy'ari, KPU members Idham Holik, Mochammad Afifudin, August Mellaz, Yulianto Sudarajat, Betty Epsilon Idroos, and Parsadaan Harahap together with KPU Secretary General Bernad Dermawan Sutrisno held a Press Conference to Determine the Candidate Pairs for President and Vice President for the 2024 Election, at KPU Media Center, Monday (13/11/2023). As Head of the Technical Implementation Division, Idham said that the KPU had

ISSN Print : 1979-7141

ISSN Online : 2541-1942

named three pairs of presidential and vice presidential candidates in the 2024 Election, namely Anis Rasyid Baswedam-Muhaimin Iskandar, Ganjar Pranowo-Mahfud MD, and Prabowo Subianto-Gibran Rakabuming Raka. The three pairs of candidates have fulfilled the provisions of article 220 of Law Number 7 of 2017 which states that political parties combining political parties can register prospective pairs of candidates, namely having fulfilled the provisions of 25% of seats in the DPR or 25% of valid votes nationally (KPU, 2023).

The video podcast uploaded on Deddy Corbuzier's YouTube channel has been watched by more than 2.6 million viewers in about 4 months. In his upload, there are hundreds or even thousands of public opinions that fill the comments column on Deddy Corbuzier's uploaded video. This of course requires a long time to group public sentiment manually. Therefore, researchers applied machine learning using the Naïve Bayes Classifier algorithm to classify sentiment related to public opinion regarding Deddy Corbuzier's upload with BEM UI regarding the ideas of presidential candidates in the 2024 elections.

The Naïve Bayes Classifier algorithm is widely used in previous research to analyze sentiment classification. According to Hamzah (2012), the Naïve Bayes Classifier has many advantages, one of which is that it is fast in calculations, a simple algorithm, and can produce high accuracy. Previous research using the Naïve Bayes Classifier was research on public sentiment towards the Independent Campus policy based on comments on YouTube using TF-IDF and Naïve Bayes weighting which produced the best accuracy of 97% (Zhafira et al., 2021). Other research, namely regarding aspects of the Female Daily review, uses TF-IDF and Naïve Bayes which produces an accuracy of 94.17% (Yutika et al., 2021). There is also other research regarding JD.ID Online Store Customer Sentiment using the Naïve Bayes Classifier based on Emotional Icon Conversion which produces an accuracy of 98% (Sari & Wibowo, 2019). Based on the explanation above, it shows that the Naïve Bayes Classifier is expected to get quite good results for sentiment analysis regarding YouTube users' comments on Deddy Corbuzier's Podcast video with BEM UI regarding the ideas of this presidential candidate.

This research applies several things including the data labeling stage using Python, text preprocessing, Term Frequency Inverse Document Frequency (TF-IDF) word weighting, then validating the data using k-fold cross validation to get not only good but also valid results.

2. RESEARCH METHODS

The algorithm used in the research is the Naïve Bayes Classifier algorithm with the correction of non-standard words at the text preprocessing stage, plus TF-IDF word weighting as represented in Figure 1. Sentiment classification is implemented in Google Colaboratory using Python and Jupyter Notebook which is an online-based service.

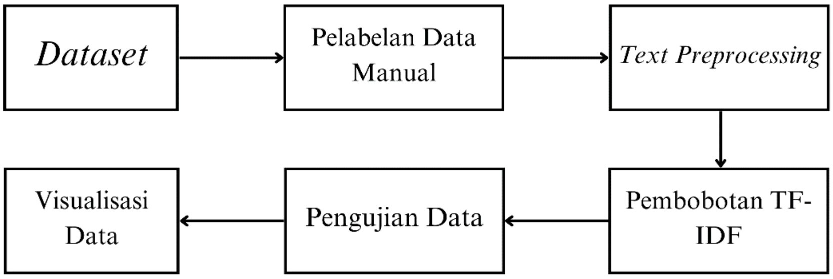


Figure 1. Research Methodology

ISSN Print : 1979-7141

ISSN Online : 2541-1942

2.1. Datasets

The data used in the research is data that is publicly available on the YouTube platform. The data collection technique used is a scraping technique. The scraping technique is a technique for getting information from a page automatically without having to copy it manually (Zhafira et al., 2021). This scraping process is carried out through Google Collaboratory using the Python language. The process of scraping data sourced from YouTube utilizes the YouTube Data API facility which is available in the resources menu. The overall dataset is 5001 comment data sorted based on the most likes.

2.2. Manual Data Labeling

The data that has been obtained is labeled manually using Google Collaboratory in Python. This process is carried out by first translating each word into English, then labeling it automatically using the Textbloob algorithm. Examples of results from data labeling can be seen in Table 1

Table 1. Example of Data Labeling

Text	Translated Text	Subjectivity	Polarity	Sentiment
Kita ga butuh pemimpin yang hanya bisa ngomong. Tapi yang tegas menentukan nasib bangsa tanpa adanya intervensi darimana pun!!!!	We don't need a leader who can only talk. But one who decisively determines the fate of the nation without any intervention from anywhere!!!	0.7	-0.1	Negatif

2.3. Text Preprocessing

Text Preprocessing is a process for normalizing terms from sentences. This is done to receive good training data and the features that are extracted will then be in sync with what is desired, thereby simplifying data processing. Collecting comment data from the YouTube platform must not be identical to standard words, words that are not in the dictionary, or regional languages that are used or omitted. To return some text to natural text by eliminating atypical expressions in order to minimize noise at a later stage, pre-processing or normalization is needed to overcome this (Gifari et al., 2022). The stages in preprocessing are as follows:

2.3.1. Case Folding and Cleansing

At this stage, all text is standardized into lowercase letters and cleaning or deleting all documents containing numbers, emoticons, characters (#*^$%), delimiters such as commas (,) and periods (.) and also signs. read others (Septian et al., 2019). Table 2 illustrates the case folding and cleaning process.

Table 2. Example of Case Folding and Cleansing results

Text	*Case Folding* dan *Cleansing*
Well done, fellow students. I personally support the student movement in this matter. Let's build this country in a better direction. Spirit! ✊	great, my fellow students, i personally support the student movement in this case, let's build this country in a better direction

2.3.2. Tokenizing

At this stage, the process of separating sentences into single words is carried out and checking words from the first character to the last character (Kosasih & Alberto, 2021). Table 3 illustrates the tokenizing process.

Table 3. Example of Tokenizing results

Case Folding dan *Cleansing*	*Tokenizing*
great, my fellow students, i personally support the student movement in this case, let's build this country in a better direction	excellent; younger brother; student; i; personal; support; movement; in; matter; this; let; get up; country; toward; which; more; good; spirit

2.3.3. Stemming

Stemming is the process of removing prefixes and suffixes in a word to get the root word from a document (Merinda Lestandy et al., 2021). In Table 4 the stemming process is illustrated.

Table 4. Example of Stemming results

Tokenizing	*Stemming*
excellent; younger brother; student; i; personal; support; movement; in; matter; this; let; get up; country; toward; which; more; good; spirit	excellent; younger brother; student; i; personal; support; movement; in; matter; this; let; get up; country; direction; which; more; good; spirit

2.3.4. Stopword Removal

Stopword removal is removing words that have no sentiment. For example, conjunctions and, with, which, and the like. Apart from that, we also remove non-conjunctive words that have no sentiment such as nouns, people's names, street names, hotel names, time pronouns, or descriptions of places (Baskoro et al., 2021). In Table 5 the stemming process is illustrated.

Table 5. Example of Stemming Results

Stemming	*Stopword Removal*
excellent; younger brother; student; i; personal; support; movement; in; matter; this; let; get up; country; direction; which; more; good; spirit	excellent; younger brother; student; personal; support; movement; let; get up; country; direction; spirit

2.4. TF-IDF weighting

The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm is a way of giving weight to the relationship of a word (term) to a document. TF-IDF is a statistical measure used to evaluate how important a word is in a document or in a group of words. In the TF-IDF algorithm, a formula is used to calculate the weight (W) of each document for keywords using the formula, namely (Silalahi & Ginting, 2022).

$$Wdt = tfdt * ldft$$

Information:
Wdt : weight of the d-document against the t-word
tfdt : the number of words searched for in a document
ldft : Inversed Document Frequency (log(N/df))
N : total dokuments
df : many documents contan the word *yang*

Table 6. Example of feature extraction results with TF-IDF

Stopword Removal	TF-IDF Extraction Results
excellent	0.213
younger brother	0.518
student	0.217
personal	0.268
support	0.251
movement	0.322
let	0.287
get up	0.287
country	0.207
direction	0.287
spirit	0.225

2.5. Data Testing

Data testing is carried out by comparing predicted data (classification results using the designed application) and actual data (classification results using manual labeling by humans). The actual data was then analyzed using the Rapidminer application with the Naïve Bayes algorithm (Aziz & Fauziah, 2022).

2.6. Data Visualization

After all stages have been carried out, the data can be visualized using the word cloud technique to display the key words that most frequently appear in YouTube user comment data on Deddy Corbuzier's video uploads with BEM UI regarding the ideas of presidential candidates. This word cloud visualization helps provide a visual overview of the most talked about topics and facilitates analysis of emerging sentiment. Visualization helps users analyze and reason about information and evidence to make complex information easier to understand and understand (Tupari et al., 2023).

3. RESULTS AND ANALYSIS

The sentiment analysis process in this research uses the RapidMiner tool version 10.2.000. The data used is YouTube user comments on Deddy Corbuzier's video podcast with BEM UI regarding the ideas of presidential candidates. The data used is 5001 (adjusting to the data limit in Rapidminer) which is sorted based on the most likes from the comments. The data that has been collected is then divided into testing data and training data, where the composition and accuracy of each division is shown in Table 7.

The model for classifying with the Naïve Bayes algorithm is as shown in the following image.

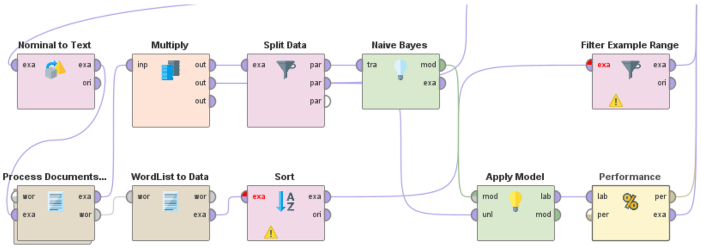


Figure 2. Naïve Bayes Algorithm Process

Testing the classification method uses Naïve Bayes by dividing the data into ten treatments with an iterative process to obtain accuracy values for each treatment which are presented in Table 7 below.

Table 7. Accuracy Results

Data Training	*Data Testing*	Accuracy
0.95	0.05	57.6%
0.9	0.1	53.65%
0.85	0.15	53.2%
0.8	0.2	53.5%
0.75	0.25	53.92%
0.7	0.3	54.27%
0.65	0.35	53.49%
0.6	0.4	53.55%
0.55	0.45	54.18%
0.5	0.5	55.88%

From the visualization of the table above, it can be seen that the highest accuracy was obtained at 57.6% with a treatment of 0.95 for training data and 0.05 for testing data. An accuracy value of 57.6% means that around 57.6% of all experimental data used was predicted correctly by the model. This shows that the Naïve Bayes model is arguably not effective enough in commenting on YouTube users' videos on Deddy Corbuzier's podcast video with BEM UI regarding the ideas of presidential candidates based on the data used to train and visualize them well.

The clean data is then labeled using the Textbloob algorithm model. With the results as in Table 8 below.

Table 8. Labeling Results

Label	Amount	Presentase
Positif	2601	52.01%
Negatif	730	14.59%
Netral	1670	33.40%

From the visualization of Table 8 above, it can be seen that, from the results of sentiment analysis of YouTube user comments on Deddy Corbuzier's video podcast with BEM UI regarding the ideas of presidential candidates using the Naïve Bayes method, there are 5001 that have been analyzed. On the positive comments, it can be concluded that almost more than fifty percent dominate. It can also be said that some are still neutral regarding the content of the video in question, this may be due to the limited information received regarding the ideas of each presidential candidate. So this can be used as a lesson to provide more information to social media users.

The data preprocessing model uses Rapidminer software as follows.

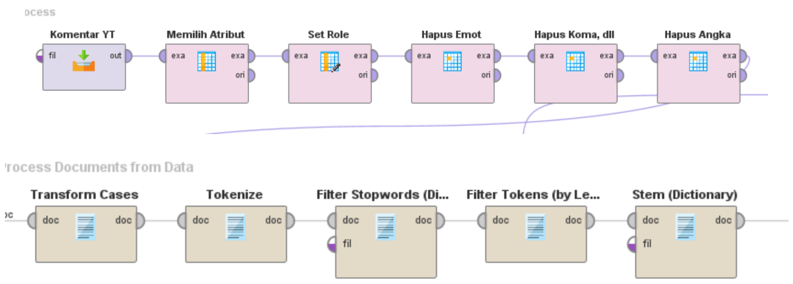
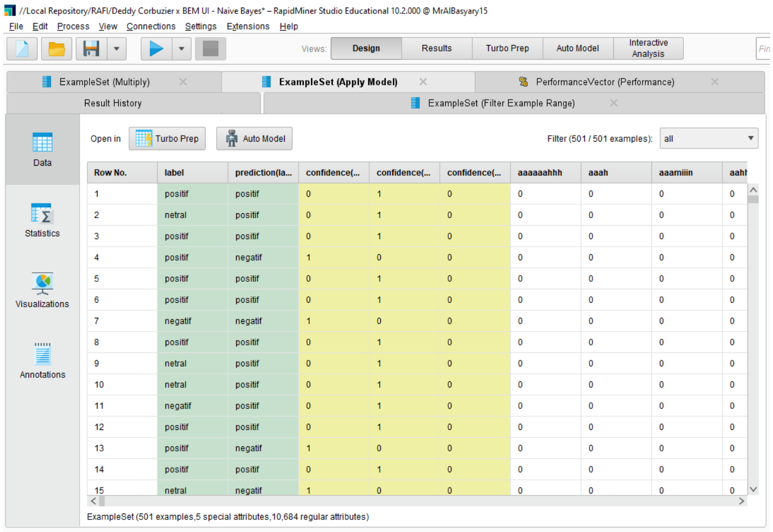


Figure 3. *Text Preprocessing*

As for the Text Preprocessing results, you can see them in Table 2 to Table 6 above..

Word weighting according to TF-IDF is carried out after preprocessing. The weighted value for each comment resulting from data processing on the attribute is then compared with each probability according to TF-IDF weighting. The results of the probabilities of positive and negative attributes are compared so that the attribute has a greater probability. If the probability of a positive opinion exceeds a negative opinion, then the document is a positive opinion and vice versa. An example of the weighting results is as shown in Figure 4 below.


//Local Repository/RAFI/Deddy Corbuzier x BEM UI - Naive Bayes* – RapidMiner Studio Educational 10.2.000 @ MrAlBasyary15
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
ExampleSet (Multiply)
ExampleSet (Apply Model)
PerformanceVector (Performance)
Result History
ExampleSet (Filter Example Range)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model

Filter (501 / 501 examples): all

Row No.	label	prediction(la...	confidence(...	confidence(...	confidence(...	aaaaaahhh	aaah	aaamiiin	aahh
1	positif	positif	0	1	0	0	0	0	0
2	netral	positif	0	1	0	0	0	0	0
3	positif	positif	0	1	0	0	0	0	0
4	positif	negatif	1	0	0	0	0	0	0
5	positif	positif	0	1	0	0	0	0	0
6	positif	positif	0	1	0	0	0	0	0
7	negatif	negatif	1	0	0	0	0	0	0
8	positif	positif	0	1	0	0	0	0	0
9	netral	positif	0	1	0	0	0	0	0
10	netral	positif	0	1	0	0	0	0	0
11	negatif	positif	0	1	0	0	0	0	0
12	positif	positif	0	1	0	0	0	0	0
13	positif	negatif	1	0	0	0	0	0	0
14	positif	positif	0	1	0	0	0	0	0
15	netral	negatif	1	0	0	0	0	0	0

ExampleSet (501 examples,5 special attributes,10,684 regular attributes)

Figure 4. Example of weighting results

Visualization of the results of sentiment analysis using the world cloud on Rapidminer was carried out to see an overview of the frequency of words that frequently appear. The purpose of this visualization is to facilitate the understanding and interpretation of sentiment analysis results visually.

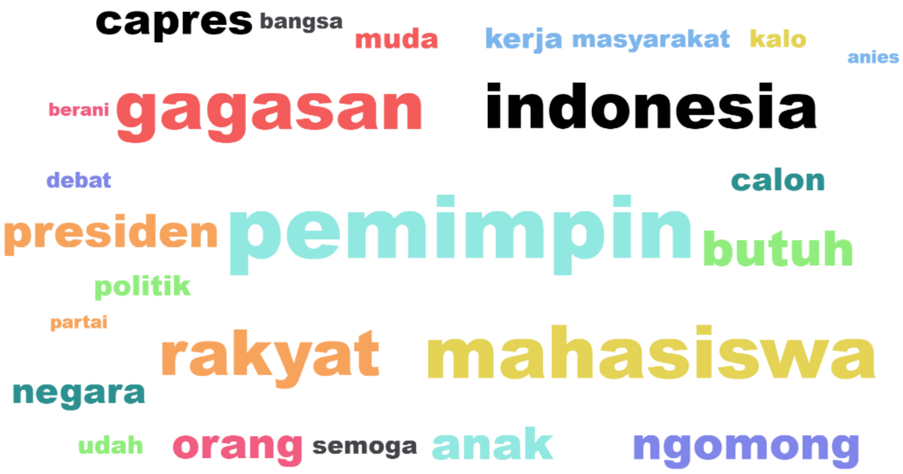


Figure 5. *Word Cloud* Sentiment

Based on the visualization image above, it can be seen that the word "Leader" dominates the word cloud, which indicates that a leader is an important aspect of the ideas desired by the Indonesian people based on the results of YouTube users' comments on Deddy Corbuzier's video podcast with BEM UI regarding ideas. presidential candidate.

4. CONCLUSION

Through the use of the Naïve Bayes algorithm, this research succeeded in identifying positive and negative sentiments from the public regarding the idea of presidential candidates for the 2024 elections, including data visualization, this information provides an overview and reactions of the public regarding this matter. The results of the Naïve Bayes algorithm show the effectiveness of using Naïve Bayes in providing sentiment analysis with an accuracy value of 57.6%, which indicates that Naïve Bayes can be said to be not effective enough in analyzing sentiment in the context of this research.

The advantage of this research is that the visualization of sentiment analysis data provides a clear and informative picture of the public's views on the ideas of presidential candidates for the 2024 election.

The weakness of this research is the limited use of algorithms without other algorithms to help improve and compare accuracy results because the accuracy obtained in this research is still quite low.

5. DECLARATION OF COMPETING INTEREST

We declare that we have no conflict of interest.

6. REFERENCES

Aziz, A., & Fauziah, F. (2022). Analisis Sentimen Identifikasi Opini Terhadap Produk, Layanan dan Kebijakan Perusahaan Menggunakan Algoritma TF-IDF dan SentiStrength. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *6*(1), 115–125.

Baskoro, B. B., Susanto, I., & Khomsah, S. (2021). Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus: Ulasan Pelanggan Pada Situs TRIPADVISOR). *INISTA (Journal of Informatics Information System Software Engineering and Applications)*, *3*(2), 21–29.

Gifari, O. I., Adha, Muh., Freddy, F., & Durrand, F. F. S. (2022). Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine. *Journal of Information Technology*, *2*(1), 36–40. https://doi.org/10.46229/jifotech.v2i1.330

Kosasih, R., & Alberto, A. (2021). Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor. *InfoTekJar: Jurnal Nasional Informatika Dan Teknologi Jaringan*, *6*(1), 134–139.

KPU. (2023, December 13). *KPU Tetapkan Tiga Pasangan Calon Presiden dan Wakil Presiden Pemilu 2024*. https://www.kpu.go.id/berita/baca/12081/kpu-tetapkan-tiga-pasangan-calon-presiden-dan-wakil-presiden-pemilu-2024#:~:text=Selaku%20Ketua%20Divisi%20Teknis%20Penyelenggaraan%2C%20Idham%20menyampaikan%2C%20KPU,Ganjar%20Pranowo-Mahfud%20MD%2C%20serta%20Prabowo%20Subianto-Gibran%20Rakabuming%20Raka.

Merinda Lestandy, Abdurrahim Abdurrahim, & Lailis Syafa'ah. (2021). Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(4), 802–808. https://doi.org/10.29207/resti.v5i4.3308

Sari, F. V., & Wibowo, A. (2019). Analisis sentimen pelanggan toko online Jd. Id menggunakan metode Naïve Bayes Classifier berbasis konversi ikon emosi. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *10*(2), 681–686.

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent System and Computation*, *1*(1), 43–49. https://doi.org/10.52985/insyst.v1i1.36

Silalahi, N., & Ginting, G. L. (2022). Analisa Sentimen Masyarakat Dalam Penggunaan Vaksin Sinovac Dengan Menerapkan Algoritma Term Frequence – Inverse Document Frequence

(TF-IDF) dan Metode Deskripsi. *Journal of Information System Research (JOSH)*, *3*(3), 206–217. https://doi.org/10.47065/josh.v3i3.1441

Tupari, T., Abdullah, S., & Chairani, C. (2023). Visualisasi Data Analisa Sentimen RUU Omnibus Law Kesehatan Menggunakan KNN dengan Software RapidMiner. *Jurnal Informatika: Jurnal Pengembangan IT*, *8*(3), 261–268. https://doi.org/10.30591/jpit.v8i3.5641

Yutika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF-IDF dan Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 422. https://doi.org/10.30865/mib.v5i2.2845

Zhafira, D. F., Rahayudi, B., & Indriati, I. (2021). Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF-IDF Berdasarkan Komentar pada Youtube. *Jurnal Sistem Informasi, Teknologi Informasi, Dan Edukasi Sistem Informasi*, *2*(1). https://doi.org/10.25126/justsi.v2i1.24