CEPAT Journal of Computer Engineering: Progress, Application and Technology
Vol. 3, No. 2, October 2024, pp. 1-8
ISSN: 2963-6728, DOI: https://doi.org/10.25124/cepat.v3i02.7734

Sentiment analysis of public comments on “program makan siang gratis” using KNN (K-nearest neighbor) and SMOTE algorithm

Vincent[1], Vincentius Hansel Irsansaputra[1], Daniel Udjulawa[1]
[1]Department of Computer Science and Engineering, School of Informatics, Multi Data Palembang University, Indonesia

Article Info

Article history:

Received June 2, 2024
Revised July 4, 2024
Accepted July 4, 2024

Keywords:

SMOTE
KNN
Sentiment

ABSTRACT

This research analyses public sentiment on “Program Makan Siang Gratis” using the KNN algorithm, enhanced with the SMOTE technique, to provide insights and recommendations for policymakers aiming to achieve “Indonesia Emas 2045”. The study employs Google Colab and Python for testing. KNN is used for sentiment analysis and classification, with SMOTE addressing data imbalance. Results from two scenarios without SMOTE and with SMOTE show that performance is more optimal without SMOTE, as SMOTE decreases performance by 34%. The k4 parameter yields the best results: 76% accuracy, 57% precision, 77% recall, and 65% F1-Score. Analysis of comments from the “tempodotco” YouTube channel reveals that public sentiment towards the program proposed by President Prabowo Subianto and Vice President Gibran Rakabuming Raka is predominantly negative.

Corresponding Author:

Vincent
Department of Computer Science and Engineering
School of Informatics, Multi Data Palembang University
Palembang, Indonesia
Email: vincent.cent@mhs.mdp.ac.id

1. INTRODUCTION

The Presidential Election is a 5-Year Agenda that is held to choose who is worthy of holding the title “Head of State”. Parties’ coalition with each other to nominate a pair of candidates that they entrust. This election is officially held by the State Official Institution of the General Election Commission or often called KPU. The election process takes a long time and a very complicated process in calculating valid votes.

The 2024 Presidential Election introduced various visions and missions from the candidates, one of which was the Prabowo Subianto-Gibran Rakabuming Raka pair with the vision of "Bersama Indonesia Maju" towards “Indonesia Emas 2045”. To achieve this goal, they refer to three major pillars: “Asta Cita” which consists of eight main missions, “Program Prioritas” with 17 major activity plans, and “Program Terbaik Cepat” which includes eight important agendas.

One of the main concerns in the “Program Terbaik Cepat” is the provision of free lunch and milk in schools as well as nutritional assistance for children under five and pregnant women. This program aims to eradicate stunting and improve the quality of human resources, with the hope of producing a healthier and smarter generation. The program is daily and targets preschool to high school students as well as Islamic boarding schools. In addition, pregnant women and children under five also receive nutritional assistance to support their health.

While the program has noble goals, public responses to its implementation have been mixed. A sentiment analysis of community comments can provide important insights into public perceptions of the

program. This analysis can help in evaluating the success of the program and identifying areas that require improvement.

In this context, the K-Nearest Neighbour (KNN) method is used to analyse the sentiment of public comments. KNN is one of the methods for machine learning that has been recognized as a method for classifying objects based on datasets used as learning data [1]. KNN is also known as a generalized algorithm for the nearest neighbour rule [2]. According to Lubis as cited by Munazhif, the KNN method is the right way to classify data based on data that is close to the desired object [3]. KNN is a simple yet effective algorithm for classification that identifies patterns in data based on proximity to other data. However, one of the challenges in sentiment analysis is data imbalance, where the number of positive and negative comments may be unbalanced.

Usually, to avoid deleting a significant majority of instances, Oversampling algorithms are preferred, and the Synthetic Minority Oversampling Technique (SMOTE) algorithm proposed by Chawla is the most widely used [4]. SMOTE is one of the methods used to cope with imbalanced data by artificial data replication for minority data classes [5]. Therefore, to overcome data imbalance in this study, the Synthetic Minority Over-sampling Technique (SMOTE) technique is used. Based on what was quoted by Nugraha from Azmatul, SMOTE increases the amount of minor class data to be equal to the major class by generating artificial data [6]. With the combination of KNN and SMOTE, sentiment analysis is expected to provide more accurate and representative results.

This research aims to analyse the sentiment of public comments on the free lunch program using the KNN algorithm enhanced with the SMOTE technique. The results of this research are expected to provide a clearer picture of public opinion and provide useful recommendations for policy makers to improve the effectiveness of the free lunch program to achieve the “Indonesia Emas 2045”.

2. METHOD

This research was conducted on March 16, 2024, the methodology in this research is divided into 2, namely the Research Stage, and Data Processing and Analysis. The following are the details of this research process:

2.1. Research Stage

This research was tested using the Google Colab website as a Cloud Server-based compiler with the Python programming language. The machine learning method approach is KNN (K-Nearest Neighbor) which is used as a model in analysing and classifying sentiment and using SMOTE (Synthetic Minority Oversampling Technique) in helping to handle data imbalance.

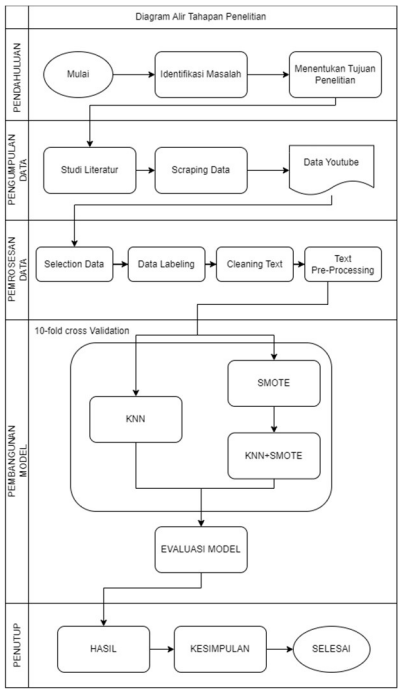


Figure 1. Research Stages

The first stage is the preliminary stage to conduct a case study by identifying the problems that will be raised in sentiment analysis. Then at the data collection stage which aims to support the sentiment analysis

research process. The data used as objects in this study are a collection of comments from the Indonesian public through the "tempotdotco" Youtube channel in a Podcast video entitled "Program Makan Siang Gratis, Anggaran Tipis dan Potensi Bagi-bagi Kue | Bocor Alus Politik" using the Netlytic website to pull and create datasets. Furthermore, after the data has been successfully pulled and exported into an Excel file in CSV (Comma Separated Values) format, data cleaning and labeling is carried out to make it more structured and easy to analyze. After the data is clean and more structured, then the next stage, which is the development of machine learning models, there are 2 scenarios in developing the model, namely the KNN model, and the KNN model which has been optimized by the SMOTE method. Finally, evaluate the results of the research and draw conclusions from the results of this study.

2.2. Data Processing and Analysis

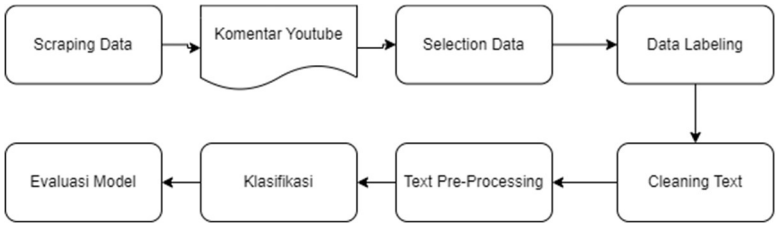


Figure 2. Data Processing and Analysis Process

The data scraping process is carried out to collect comment data on the "tempodotco" Youtube channel on a video entitled "Program Makan Siang Gratis, Anggaran Tipis dan Potensi Bagi-bagi Kue | Bocor Alus Politik" with a total of around 2,500 comments.

From the scraping results, selection is made on datasets that do not match the research criteria, such as comments containing "SARA" (Ethnicity, Religion, and Race), duplicate comments, online gambling promotion comments, and so on that are not related at all to the theme and title of the video. After that, manually label the data into neutral, positive, and negative classes.

The next stage is text *pre-processing* which consists of:

1. *Casefolding* is the process of converting all letters in a document or sentence into lowercase letters [7]. *Casefolding* is used to make searching easier. Not all data is consistent in the use of capital letters.
2. *Tokenizing* which is the process of separating or breaking a sentence in a document into separate words (tokens) based on characters, spaces, and maybe at the same time the process of removing certain characters, such as punctuation marks [7].
3. *Stopword Removal* which is the process of removing words that are not needed or have no meaning such as conjunctions, pronouns, and others [7].
4. *Stemming* is a process used to change the *terms* that are still attached to the term prefixes, inserts, and suffixes. The *stemming* process is done by removing all *affixes* consisting of *prefixes*, *infixes, suffixes*, and *confixes* (a combination of prefixes and suffixes) [8].

Furthermore, after passing the text *pre-processing* stage, determine the classification by using the *sklearn* CountVectorizer *library* to convert text into numeric for easy *machine learning* model development. Then the last stage, evaluate the performance of the model. *Confusion matrix* is the most popular tool in evaluating classification performance. The following table shows the *confusion matrix* for binary classes [9].

Table 1. Confusion Matrix of Binary class

Class	Predictive Positive	Predictive Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

From the values contained in the *confusion matrix*, it can then produce values used to evaluate the classification method, namely the *accuracy, precision, recall,* and *f1-score* values. The accuracy, precision, recall, and f1-score values can be calculated through equations (5-8) [9], [10]

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{FN + TP} \quad (7)$$

$$F1 - score = \frac{2TP}{2TP + FN + FP} \quad (8)$$

3. RESULTS AND DISCUSSION

The implementation of the research stages encompasses a comprehensive process beginning with the collection of YouTube comment data and culminating in the evaluation of the model using a confusion matrix.

3.1. Results

At this stage, the implementation of the research stages will be carried out, starting from collecting *Youtube* comment data to evaluating the model using *confusion matrix.*

1) Data Collection

Data collection was carried out by *scraping* the comments column on the "tempodotco" Youtube channel on the video entitled "Program Makan Siang Gratis, Anggaran Tipis dan Potensi Bagi-bagi Kue | Bocor Alus Politik" using the *Netlytic* website in its collection, then obtained as many as 2,500 comments.

2) Data Selection

After the data *scraping* process and getting the results of 2,500 comments, the next selection process was carried out so that the number of comments to be used in the study from 2,500 to 1000 comments.

3) Data Labelling

The data selection process has been completed, then labelling the data with negative, positive, and neutral sentiments. Labelling is done manually with negative sentiment comment data by 20.1%, positive by 4.2%, and neutral by 75.7%.

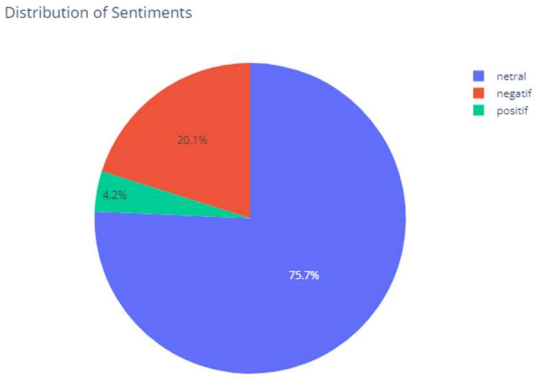


Figure 3. Sentiment Distribution

4) Data Cleaning

The next stage is data cleaning, starting from changing abbreviations such as egk, g, etc. into sentences that should be written and standardized such as egk to no, punctuation, and emoji removal.

5) Text *pre-processing*

The text pre-processing process consists of case folding, cleansing, tokenizing, normalizing, stopword removal, and stemming.

a) Case folding

Here are the results of the *casefolding* process

Before	Percuma makan siang gratis tapi biaya pendidikan mahal
After	percuma makan siang gratis tapi biaya pendidikan mahal.

b) *Cleansing*

Here are the results of the *cleansing* process

Before	Makan siang gratis dan minum susu yg mengandung asam sulfat 😅
After	makan siang gratis minum susu mengandung asam sulfat

c) *Tokenizing*

Here are the results of the *tokenizing* process

Before	percuma makan siang gratis tapi biaya pendidikan mahal.
After	[percuma, makan, siang, gratis, tapi, biaya, pendidikan, mahal]

d) *Stopword Removal*

Here are the results of the *stopword removal* process.

Before	[percuma, makan, siang, gratis, tapi, biaya, pendidikan, mahal]
After	[percuma, makan, siang, gratis, biaya, pendidikan, mahal]

e) *Stemming*

Here are the results of the *stemming* process

Before	[dana, diambil, ukt, mahasiswa]
After	dana ambil ukt mahasiswa

6) Classification

Each data modeling scenario will be formed using the *10-fold cross validations* technique. [10]. At this stage, *cross validations are* carried out. By testing the *k* parameter in KNN (K-Nearest Neighbor) *machine learning* modeling in the range of *k = 3* to *k = 15* and taking only negative and positive sentiment values without neutral sentiment values.

The following are the results of testing the parameters k = 3 to *k = 15* in KNN modeling, the results are as follows:

Table 2. KNN performance results

K	Accuracy	Precision	Recall	F1-Score
3	0.68	0.61	0.68	0.64
4	0.76	0.57	0.77	0.65
5	0.70	0.56	0.70	0.62
6	0.75	0.57	0.76	0.65
7	0.76	0.57	0.76	0.65
8	0.76	0.57	0.76	0.65
9	0.76	0.57	0.76	0.65
10	0.76	0.57	0.76	0.65
11	0.76	0.57	0.76	0.65
12	0.76	0.57	0.76	0.65
13	0.76	0.57	0.76	0.65
14	0.76	0.57	0.76	0.65
15	0.76	0.57	0.76	0.65

Judging from the accuracy value produced by each KNN parameter in table 2.7, *k = 4* is the parameter with the most optimal value. The *accuracy* value is 76%. Next, the *confusion matrix* value is calculated on the optimal parameter *k = 4* shown in Figure 4.

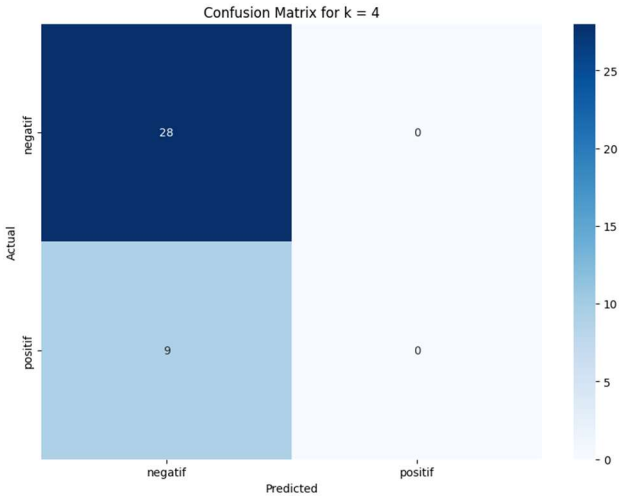
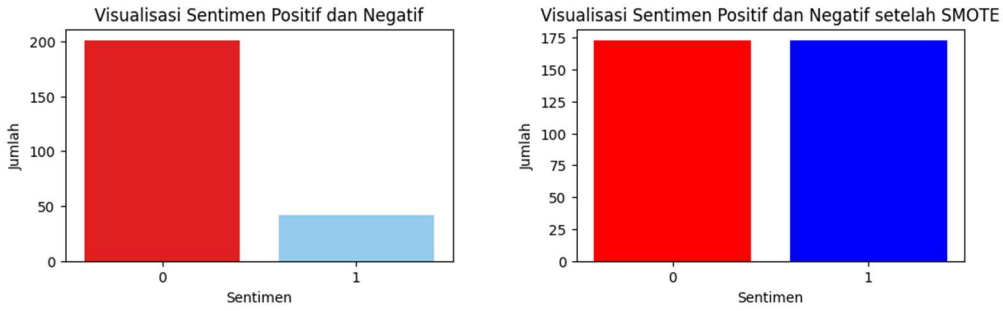


Figure 4. KNN *Confusion Matrix*

Furthermore, the second model scenario is KNN optimization by applying the SMOTE method to balance the distribution of sentiment classes.



(a) (b)

Figure 5. Visualization of positive and negative Sentiments: (a) before SMOTE, (b) after SMOTE

When looking at the comparison of the distribution of positive and negative sentiments before SMOTE, the sentiment is not balanced between negative (0) and positive (1), and after SMOTE the sentiment data has been balanced with a total of 175 data. After the data is balanced, the next step is to start the KNN optimization model training process with SMOTE.

Table 3. Performance results of KNN with SMOTE

K	Accuracy	Precision	Recall	F1-Score
3	0.22	0.35	0.22	0.16
4	0.24	0.43	0.24	0.18
5	0.24	0.43	0.24	0.14
6	0.24	0.43	0.24	0.14
7	0.22	0.05	0.22	0.09
8	0.22	0.05	0.22	0.09
9	0.24	0.06	0.24	0.1
10	0.24	0.06	0.24	0.1
11	0.24	0.06	0.06	0.24
12	0.24	0.06	0.24	0.1
13	0.24	0.06	0.24	0.1
14	0.24	0.06	0.24	0.1
15	0.24	0.06	0.24	0.1

Judging from the accuracy value produced by each KNN parameter in table 2.8, the parameter that has the optimal value in implementing SMOTE is $k = 4$ with an *accuracy* value of 24%. Next, the *confusion matrix* value is calculated on the optimal parameter $k=4$ which is shown in Figure 3.5.

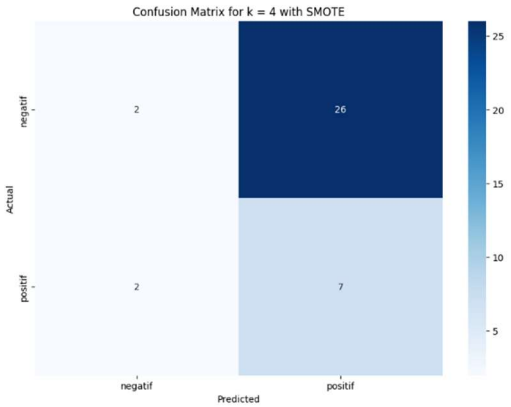


Figure 6. Confusion Matrix KNN with SMOTE

3.2. Discussion

Based on the classification model training process, two model training scenarios were conducted due to data imbalance. Likely to affect the performance and performance given by the model that has been trained, the SMOTE method is used to balance the data. After performing the model classification stage, a comparison of the KNN model without SMOTE and with SMOTE is obtained, the following are the results

Table 4. Model Performance Comparison

Performance	KNN	SMOTE	Improved
Accuracy	76%	24%	-52%
Precision	57%	43%	-14%
Recall	77%	24%	-53%
F1-Score	65%	18%	-47%
Average			-34 %

Table 4 shows that there is a decrease in model performance after the second scenario, namely using the SMOTE method. The *accuracy* value which was originally 76% to 24% decreased very drastically by 52%. The *precision* value decreased by 14% from 57% to 43%. *Recall* value initially 77% to 24% decreased by 53%, and the last *F1-Score* decreased by 47% from 65% to 18%. With the comparison results of the two scenarios, it can be concluded that the second scenario, which uses the SMOTE method, is not effective because it makes the model performance decrease by an average of 34%.

Based on the comments of the tempodotco Youtube channel from the video entitled "Program Makan Siang Gratis, Anggaran Tipis dan Potensi Bagi-bagi Kue | Bocor Alus Politik", here is a visualization of sentiment using the *wordcloud library*.

(a) (b)

Figure 7. Sentiment visualization: (a) negative, (b) positive

Overall, the responses in the comment section of tempodotco's Youtube channel showed that negative sentiment had the highest number of 200 comments about the free lunch program proposed by Mr. Prabowo Subianto as President and Gibran Rakabuming Raka as Vice President.

4. CONCLUSION

Based on the results of model classification with two scenarios, namely without using the SMOTE method and using the SMOTE method, it is evident that performance tends to be more optimal without the SMOTE method while the use of SMOTE reduces modelling performance very drastically by 34% in average. The k=4 parameter is the most optimal parameter with an accuracy value of 76%, a precision value of 57%, a recall value of 77%, and finally F1-Score with a value of 65%. From the public comments on tempodotco's Youtube channel, it can be concluded that people tend to have negative sentiments towards the free lunch program proposed by President Prabowo Subianto and Vice President Gibran Rakabuming Raka.

REFERENCES

[1] A. Pamuji, "Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment," *Juisi*, vol. 07, no. 01, 2021.
[2] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2019.2955754.
[3] N. F. Munazhif, G. J. Yanris, and M. N. S. Hasibuan, "Implementation of the K-Nearest Neighbor (kNN) Method to Determine Outstanding Student Classes," *SinkrOn*, vol. 8, no. 2, 2023, doi: 10.33395/sinkron.v8i2.12227.
[4] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Applied Nanoscience (Switzerland)*, vol. 13, no. 3, 2023, doi: 10.1007/s13204-021-02063-4.
[5] R. D. Permatasari, W. Rizki, and N. N. Debataraja, "Penerapan synthetic minority oversampling technique dalam mengatasi data tidak seimbang pada metode classification and regression tree," 2020.
[6] W. Nugraha, D. Risdiansyah, D. Purwaningtias, T. Hidayatulloh, and S. Suhada, "Kombinasi Tomek-Link Dan Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Credit Card Fraud," *Jurnal Larik: Ladang Artikel Ilmu Komputer*, vol. 2, no. 2, 2022, doi: 10.31294/larik.v2i2.1789.
[7] B. Gunawan, H. Sasty, and E. Esyudha, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)*, vol. 4, no. 2, pp. 17–29, 2018.
[8] F. Amin, "Sistem Temu Kembali Informasi dengan Metode Vector Space Model," *Jurnal Sistem Informasi Bisnis*. p. 2, 2012.
[9] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE Dan K-Nearest Neighbor," *Journal Information System Development (ISD)*, vol. 3, no. 1, 2018.
[10] K. Pramayasa, I. M. D. Maysanjaya, and I. G. A. A. D. Indradewi, "Analisis Sentimen Program Mbkm Pada Media Sosial Twitter Menggunakan KNN Dan SMOTE," *SINTECH (Science and Information Technology) Journal*, vol. 6, no. 2, 2023, doi: 10.31598/sintechjournal.v6i2.1372.