Hafiidh Khoiru Pradana
Informatics Program,
Universitas Amikom Yogyakarta,
Yogyakarta, Indonesia
email:
hafiidh.pradana@students.amikom.ac.id

Yuli Astuti[1]
Informatics Program,
Universitas Amikom Yogyakarta,
Yogyakarta, Indonesia
email: yuli@amikom.ac.id

Dewi Anisa Istiqomah
Informatics Program,
Universitas Amikom Yogyakarta,
Yogyakarta, Indonesia
email: dewianisaist@amikom.ac.id

Supriatin
Informatics Program,
Universitas Amikom Yogyakarta,
Yogyakarta, Indonesia
email: supriatin@amikom.ac.id

Ninik Tri Hartantie
Informatics Program,
Universitas Amikom Yogyakarta,
Yogyakarta, Indonesia
email: ninik.t@amikom.ac.id

Twitter Sentiment Analysis Classification to Assess Public Opinion on Football Matches Using the Naïve Bayes Method

The Kanjuruhan tragedy has attracted many comments on various social media platforms. This research will compare the number of positive and negative comments on Twitter and social media and determine the accuracy of the classification method used. The data used in this study consisted of 2052 pieces, consisting of 1015 positive and 1037 negative pieces. To determine the effect of the amount of training data on the resulting accuracy, testing will be carried out three times with different combinations of training data and test data, namely 70:30, 80:20, and 90:10. The results of this study obtained the highest accuracy value of 79.6%. This program can be developed for other social media platforms such as Facebook, Instagram, and others.

***KeyWords**: Sentiment Analysis; Kanjuruhan Tragedy; Naïve Bayes; Football*

This Article was:
submited: 16-11-23
accepted: 08-12-23
publish on: 20-12-23

How to Cite:
Y. Astuti, et al, "Template for JISS Journal Papers: JISS UTDI No More Then 12 Words", Journal of Intelligent Software Systems, Vol.2, No.2, 2023, pp.37–41, 10.26798/jiss.v2i2.1136

1 Introduction

Social media is the media for interaction that is most often used by various groups of people. Various public opinions are increasingly developing and are free to be expressed on social media. Public opinion arises due to an event that occurred in Indonesia. People will interact with each other to respond to an event that occurs and will cause different opinions. The more people who discuss a topic, the topic will become a trending topic on Twitter. One of the topics that became a trending topic was #kanjuruhan. The events in Kanjuruhan became a trending topic on Twitter on October 2, 2022. The incident at Kanjuruhan occurred during a soccer match in East Java which brought Arema FC against Persebaya Surabaya (1/10/2022) night. Reporting from the detik.com page, the number of victims of the Kanjuruhan Tragedy became 737 people as of October 12, 2022. The data includes 579 victims of minor injuries, 26 seriously injured, 132 dead and 12 people still undergoing treatment at the hospital. From such a large number of victims, it is not surprising that the incident has become a topic that is widely discussed in various media. Sentiment analysis is a field of study that analyzes opinions or opinions about an entity[1].

Many sentiment analysis studies have been conducted to classify public opinion based on positive or negative classes. Sentiment analysis on the performance of the Indonesian national soccer team using the naïve bayes method has been done but has a different level of accuracy, namely 87% by Ningrum and 83% accuracy by Prajamukti[2],[3]. Meanwhile, public opinion on the Indonesian National Team at the 2020 AFF Cup Championship by utilizing a larger dataset of 10,000 data with the same method is claimed to have a high level of accuracy in classification. The accuracy result obtained is 64.41%[4]. When compared to previous research with a smaller data set of 530 data, it produces a greater accuracy value of 87%. The difference is because the process of dividing training data and testing data is different.

This research aims to determine the number of positive comments, negative comments and the accuracy value of the Naïve Bayes method on events that occurred at Kanjuruhan Malang. This is done to provide information to PSSI, PT LIB (Liga Indonesia Baru), Security, Organizing Committee, and Broadcasting Party so that in the future an evaluation of policies that must be carried out so that similar incidents do not occur again.

2 Research Methods

2.1 Research Stages. This research uses data collected from twitter social media using the keyword #kanjuruhan. The keyword #kanjuruhan was chosen because there was a tragedy at the Kanjuruhan Stadium in Malang which occurred on October 1, 2022 and caused many casualties. The tragedy caught the attention of Indonesians and even foreigners. Many comments discussed the tragedy on twitter social media and even became a trending topic

[1]Corresponding Author.

in Indonesia. The research stages used in the research of Sentiment Classification Analysis of Public Opinion on Football Matches can be divided into four stages, namely data collection, data preprocessing, modeling, and evaluation. The flow of research stages can be seen in Figure 1.

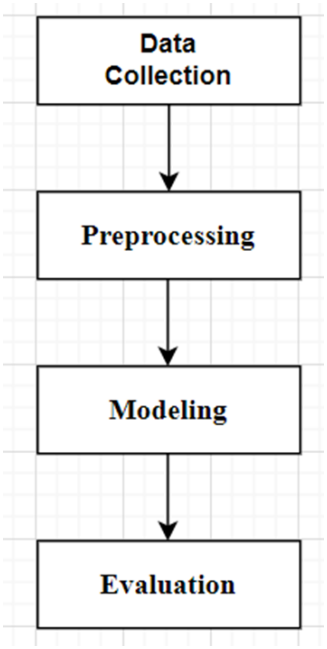

Fig. 1 Research Flow

2.2 Data Collection. The data used in this study are tweets that contain the keyword #kanjuruhan. The data collection stage in this study is to use the snscrape library. The tweets data taken from October 2, 2022 to October 12, 2022.

2.3 Pre-processing. At the data preprocessing stage, case folding will be carried out, namely to change the letters in the tweet data to lowercase letters and remove unnecessary punctuation and numbers, case folding will also remove the url address in the sentence. The next stage is Normalization which is done to correct words that are written not in accordance with the correct spelling. The next stage is stopword, which removes common words that often appear that have no meaning. The next stage is stemming, which converts each word into a base word. The last stage is Tokenizing, which separates the tweet data into individual words. By doing data pre-processing, it is hoped that it can improve the data structure so that it is ready to be processed.

2.4 Modeling. At the modeling stage, weighting is carried out on each word in the tweet data using the TF-IDF (Term Frequency Inverse Document Frequency) method. After weighting, then divide the data into two, namely training data and test data. Data division or split data is done three times with a combination of training data comparison of test funds sequentially, namely 70:30, 80:20, and 90:10. After dividing the data, the last stage is to apply classification using the Naïve Bayes method to the tweet data. The Naïve Bayes method is used because it is simple and effective when compared to other methods. The Naïve Bayes method classifies data based on probability.

2.5 Evaluation. In the evaluation and testing stage, the accuracy of the Naïve Bayes method is tested using the confusion matrix method. After the testing process then determines the number of comparisons of negative sentiment and positive sentiment so that a conclusion can be drawn.

2.6 Python. Python is an interpreted programming language that is easy to learn, focuses on code readability, and is versatile. Python is also a high-level programming language developed by Guido van Rossum. Python is widely used to create various programs such as data science, machine learning, CLI programs, GUI programs, mobile applications, web applications, Internet of Things (IoT), games, hacking programs, etc. Python is also considered as an easy-to-learn programming language as its syntax structure is clear and easy to understand[5].

2.7 Google Colab. Google Colab is an IDE for Python programming, with processing done on Google servers using powerful hardware. In software, Google Colab provides most of the required libraries[6].

2.8 Documentation. At the documentation stage, consultation is carried out on the writing of this research every process of work. The purpose of the consultation is so that every stage that has been carried out can be documented and can be useful as a reference in academic purposes. This research can also be developed based on the results and suggestions that have been given.

3 Results and Discussion

3.1 Data Collection. This research uses data obtained from Twitter social media. The data collection process is done using the Snscrape library. The data needed in this research is tweet data containing the word #kanjuruhan. In this process, more than 3000 data will be collected from October 2, 2022 to October 14, 2022.

	Tweet Id	Text	Username
0	1580709874920345602	Kepala Divisi Humas Polri, Dedi Prasetyo memas...	newsmerahputih
1	1580708875887525888	Menyalahkan Tuhan. 😡😡😡 #pssi #Kanjuruhan https...	adibjauhari
2	1580703000305487872	Dari garis Ibu, Ken Arok masih berketurunan da...	adhiyanblog
3	1580702081903886337	Tapi masyarakat Jawa yang masih memegang tradi...	adhiyanblog
4	1580667104251895808	Maaf bisa dengan mudah diberikan begitu saja/n...	ikaaks
...	...	...	...
3103	1577140196864245761	#StopPoliceBrutality #EndPoliceBrutality #Pray...	MCAOps
3104	1577139763148038144	Jika doamu tidak diberi secara instan, tetapla...	IyossWazzo
3105	1577139475850723328	A moment of silence ahead of the Real Madrid V...	peepsid
3106	1577138988799885317	Editorial	While the blame for the deaths in ...	TheHinduComment
3107	1577138548964470784	Pembantaian anak2 di #kanjuruhan !!	Adra_silent

3108 rows × 3 columns

Fig. 2 Data Collection Result

3.2 Pre-Processing. After getting the data needed, the Preprocessing process is carried out. The following are the steps in preprocessing the data.

3.2.1 Case Folding. Case Folding process is done to remove data from numbers, URLs, hashtags, delimiters such as commas and periods, and other punctuation marks. Case Folding also serves to remove sentences with emoticons, as shown in Table 1.

Table 1 Case Folding

Before	After
Trimakasih Pak Polisi (Brimob) tlah melindungi kami... #gasairmata #kanjuruhan #Indonesia #arema #persebaya https://t.co/nXYMncUNvv	trimakasih pak polisi brimob tlah melindungi kami

3.2.2 Normalization. The normalization process is carried out to correct words whose writing is not appropriate and change words in the form of abbreviations. The normalization process is carried out using a dictionary file containing incorrect words and their corrections, as shown in Table 2.

Table 2 Normalization

Before	After
trimakasih pak polisi brimob tlah melindungi kami	terima kasih pak polisi brigade mobil telah melindungi kami

3.2.3 Stop Words. The stopword process is done to remove words that are considered unimportant such as the word "yang","di","dan","ke", and others. The stopword process is done using the stopword dictionary available in the nltk library, as shown in Table 3.

Table 3 Stopword

Before	After
terima kasih pak polisi brigade mobil telah melindungi kami	terima kasih polisi brigade mobil melindungi

3.2.4 Stemming. The stemming process is done using the Sastrawi library. This process is done to convert all words into basic words, as shown in Table 4

Table 4 Stemming

Before	After
terima kasih pak polisi brigade mobil telah melindungi kami	terima kasih polisi brigade mobil lindung

3.2.5 Tokenize. In the tokenizing process is to separate the sentence into individual words. This process utilizes the nltk (Natural Language Toolkit) library by using the word_tokenize() function provided, as shown in Table 5.

Table 5 Tokenize

Before	After
terima kasih pak polisi brigade mobil telah melindungi kami	[terima, kasih, polisi, brigade, mobil, lindung]

3.3 Labeling. In this research, the VADER (Valence Aware Dictionary and Sentiment Reasoner) library is used to perform automatic labeling. The use of the VADER library must translate sentences from Indonesian to English using the Googletrans library.

3.4 Modeling. The modeling stage is carried out by calculating the weighting using the TF-IDF method. Then after that, the data is divided into training data and test data. The last stage is the application of the Naïve Bayes method.

3.5 TF-IDF Weighting. TF-IDF is one of the methods of weighting a word in an information retrieval system. TF-IDF is a method that combines Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency is the frequency of occurrence of the same word in the document. Inverse Document Frequency is the number of collections of documents that contain certain words[7]. The Term Frequency equation can be seen in formula 2. The Inverse Document Frequency equation can be seen in formula 3. The Term Frequency - Inverse Document Frequency (TF-IDF) equation can be seen in formula 4. The results of TF-IDF can be seen in Table 6.

$$W(t,d) = TF(t,d) \tag{1}$$

$$IDF(t) = log(\frac{N}{df(t)}) \tag{2}$$

$$TF - IDF = TF(t,d) \times IDF(t) \tag{3}$$

With TF(t,d) is the frequency of occurrence of term t in document d, N is the number of documents, df(t) is the number of documents containing term t[7].TF is the number of occurrences of a term. IDF is the inverse value of the frequency of words that appear in the document.

Table 6 TF-IDF Weighting Results

Term	TF-IDF			
	D1	D2	D3	D4
kandungan	0.698	0	0	0
gas	0.397	0.397	0	0
air	0.397	0.397	0	0
mata	0.397	0.397	0	0
tragedi	0.397	0	0	0.397
kanjuruhan	0.221	0	0.221	0.221
timbul	0	0.698	0	0
panik	0	0.698	0	0
konsentrasi	0	0.698	0	0
massa	0	0.698	0	0
pintu	0	0.698	0	0
mati	0	0.698	0	0
tambah	0	0	0.397	0.397
korban	0	0	0.397	0.397
insiden	0	0	0.698	0
orang	0	0	0.698	0

3.6 Data Division. In the process of dividing data into training data and test data utilizing the sklearn library and using the train_test_split module. In this research, the division was carried out three times with a combination of training data and test data comparisons sequentially 70:30, 80:20, and 90:10, as shown in Table 7

Table 7 Division of Training and Test Data

Split Data	Test Data	Training Data
70:30	616	1436
80:20	411	1641
90:10	206	1846

3.7 Naive Bayes Algorithm. Naïve Bayes Classifier (NBC) is a learning method with a simple probability concept[**?**] used to predict sentiment patterns[8] Classification using the Naïve Bayes method is carried out on training data and test data. The classification process on the training data will later be used as a model in classifying the test data. The calculation equation for the Naïve Bayes method can be seen in the formula 4.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{4}$$

Where X is data with an unknown class, H is the hypothesis that data X is a specific class, P(H|X) is the probability of hypothesis H based on condition X (a posteriori probability), P(X|H) is the probability of hypothesis X based on condition H, P(H) is the probability of hypothesis H (prior probability), and P(X) is the probability of X (prior probability) X.

3.8 Evaluation. In the testing stage, testing is carried out using the confusion matrix. Testing is done to determine the performance of the Naïve Bayes method in performing sentiment classification. Testing is done three times. The greater the accuracy value, it shows that the method is suitable and runs well in classification. In three tests, the largest accuracy value will be sought. The results are shown in Table 8, Table 10, and Table 12. Accuracies, Precisions, Recalls, and F1-scores are shown in Table 9, Table 11, and Table 13 for each confusion matrix.

Table 8 Confusion Matrix Of 70:30 Comparison

	True Positive	True Negative
Pred. Positive	252	56
Pred. Negative	77	231

Table 9 Testing with 70:30 ratio

	Positive	Negative
Precision	80%	77%
Recall	75%	82%
F1-Score	78%	79%
Accuracy	78.4%	

Table 10 Confusion Matrix Of 80:20 Comparison

	True Positive	True Negative
Pred. Positive	169	39
Pred. Negative	50	153

Table 11 Testing with 80:20 ratio

	Positive	Negative
Precision	80%	75%
Recall	75%	99%
F1-Score	77%	85%
Accuracy	78.3%	

Table 12 Confusion Matrix Of 90:10 Comparison

	True Positive	True Negative
Pred. Positive	89	18
Pred. Negative	24	75

Table 13 Testing with 90:10 ratio

	Positive	Negative
Precision	81%	79%
Recall	76%	83%
F1-Score	78%	81%
Accuracy	79.6%	

In the three tests using the confusion matrix method, the largest value was obtained with an accuracy value of 79.6% in the comparison of the division of training data and test data of 90:10 with the confusion matrix value as follows:

(1) The actual positive data results that are predicted to be positive are 89 data.
(2) Positive actual data results that are predicted to be negative are 18 data.
(3) Negative actual data results that are predicted to be positive are 24 data.
(4) The negative predicted negative actual data results are 75 data.

3.9 Conclusions. Based on the results of the research "Twitter Sentiment Classification to Assess Public Opinion on Football Matches Using the Naïve Bayes Method" can be concluded as follows: From 2052 data with the topic of discussion of the Kanjuruhan tragedy after going through preprocessing, manual labeling, weighting, and implementation, the number of positive comments is 1015 data and negative comments are 1037 data. After three tests on the dataset with manual labeling, the Naïve Bayes method gets the highest value with an accuracy value of 79.6%.
The results of the research conducted by the author are far from perfect, therefore suggestions are needed to develop this research in the future. Suggestions for this research are as follows: It is hoped that future researchers will be able to use a larger number of datasets, from this the resulting accuracy results will also be different. It is expected that future researchers will be able to use different methods. The use of different methods will produce different accuracy. It is expected in future research to be more thorough and pay attention to the process of labeling data if done manually. Labeling greatly affects the performance of the classification method.

References

[1] Septian, J. A., Fachrudin, T. M., and 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), pp. 43–49.

[2] Astiningrum, M., Haniah, M., and rahmat yoga Pradana, Y., 2020, "Analisis Sentimen Tentang Opini Terhadap Performa Timnas Sepak Bola Indonesia Pada Twitter," Seminar Informatika Aplikatif Polinema (Siap), pp. 35– 39.

[3] Prajamukti, Jayanta, d. S., 2021, "Klasifikasi Dan Analisis Sentimen Pada Data Twitter Menggunakan Algoritma Naïve Bayes (Studi Kasus: Timnas Indonesia Senior, U-23, Dan U-19)," Seinasi-Kesi, pp. 1–8.

[4] Hadiyan Sidik, M., Widiyanesti, S., and Puteri Ramadhani, D., 2022, "Analisis Sentimen dan Topic Modelling Terhadap Tim Nasional Indonesia di Kejuaraan AFF Suzuki Cup 2020 Berdasarkan Opini Pengguna Twitter Analysis of Sentiment and Topic Modeling of the Indonesian National Team in the 2020 AFF Suzuki Cup Championship Base," **9**(5), pp. 2783–2796.

[5] Melinda, R. N., Ningrum, L. M., Suryabrata, I. B., Dwipa, G. S. B. A., and Sukoco, T. P., 2021, "Program Perhitungan RAB Pekerjaan Struktur Baja (WF BEAM) Menggunakan Bahasa Python," TIERS Information Technology Journal, **2**(1).

[6] Guntara, R. G., 2023, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7 Jurnal Teknologi Dan Sistem Informasi Bisnis," Jurnal Teknologi Dan Sistem Informasi Bisnis, **5**(1), pp. 55–60.

[7] Asshiddiqi, M. F. and Lhaksmana, K. M., 2020, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI," e-Proceeding of Engineering, **7**(3), pp. 9936–9948.

[8] Setyaningsih, E. R., 2021, "Sentiment Classification untuk Opini Berita SepakBola," Journal of Intelligent System and Computation, **3**(2), pp. 93–98.