Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 7, No. 2, September 2024, pp. 541 – 552
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Sentiment Analysis Towards the Film Dirty Vote on Twitter Social Media Using the K-Nearest Neighbor Algorithm

[1]Annisa Fadillah, [2]Sriani
[1,2]Departement of Computer Science, Faculty of Science and Technology,
State Islamic University of North Sumatra, Indonesia
Email: [1]annisafadillah388@gmail.com, [2]sriani@uinsu.ac.id

Article Info

Article history:
Received Jul 10th, 2024
Revised Aug 28th, 2024
Accepted Sep 6th, 2024

Keyword:
Dirty Vote
K-Fold Cross Validation
K-Nearest Neighbor
Sentiment Analysis
Social Media

ABSTRACT

The appearance of the dirty vote film has received public attention and went viral on social media after being released and watched by millions of people in a short time. The dirty vote film has become a topic of discussion, one of which is on the social media platform Twitter. This research was conducted to determine the views or tendencies of public opinion regarding dirty vote films on Twitter social media using K-Nearest Neighbor which will be classified into positive, neutral and negative sentiment. The sentiment data that was collected in the data crawling process was 4000 pieces of data. Then after preprocessing there were 3978 data. Labeling was carried out using text blob, it was found that the negative sentiment class was 3470 superior to the positive sentiment class of 451 and the neutral sentiment class was 57. The 10-fold cross validation test produced an average accuracy value of 87.5%. Testing was carried out with 80% training data consisting of 3182 data and 20% test data consisting of 796 test data. The results of sentiment analysis show that the K-Nearest Neighbor method can be used for sentiment analysis. The accuracy value obtained was 87%, precision was 87%, recall was 100%, and f1-score was 93%.

Corresponding Author:
Annisa Fadillah,
Departement of Computer Science,
State Islamic University of North Sumatra,
Jl. Lap. Golf No.120, Kp. Tengah, Kec. Pancur Batu, Deli Serdang Regency, North Sumatra 20353.
Email: annisafadillah388@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2.32471

1. INTRODUCTION

The digital world has experienced very rapid changes and developments, resulting in the introduction of many new technologies that make it easier for people to access information and communication. The internet is one proof of technological development which is currently having an impact on society [1]. Based on data obtained from the Indonesian Internet Service Providers Association (APJII) in February 2024, the number of internet users in Indonesia in 2024 reached 221,563,479 people from a total population of 278,696,200 Indonesians in 2023. So the results of the 2024 Indonesian penetration survey released by APJII touched 79.5%, this number increased 1.31% compared to the previous period. This proves that internet users are increasing every year. The increasing use of the internet will give rise to developments in the use of platforms including social media. We Are Social recorded that social media users in Indonesia in January 2024 were 139 million social media users. According to the latest We Are Social report, there are 10 social media that are most widely used in Indonesia, including Whatsapp, Instagram, Facebook, Tiktok, Telegram, Twitter and Facebook Messenger, Pinterest, Kuaishou, LinkedIn.

Twitter is the fastest social media in disseminating information. Twitter is one of the social media that often becomes a trending issue nationally and internationally which is used by social media users as a medium to voice opinions about something that is currently being widely discussed on very complex social networks [2]. Based on data from We Are Social, in October 2023 there were 666.2 million Twitter users worldwide.

Indonesia is ranked 4th in the world with 25.25 million Twitter users as of July 2023. Currently, the use of Twitter, especially in Indonesia, has had a big impact on opinions on certain topics [3]. One of the trending topics that has recently been widely discussed on Twitter in the lead up to the 2024 election is the documentary film Dirty Vote which presents various kinds of comments and creates controversy among social media users, one of which is on Twitter.

The film dirty vote is a documentary film that depicts allegations of fraud in the election. This film can be watched on a YouTube channel called Dirty Vote. News about the launch of this film can also be found in CNN Indonesia articles, as well as public discussions and reactions on social media, especially on Twitter. The Dirty Vote film was released on February 11, 2024, directed by Dandhy Dwi Laksono and featuring several constitutional law experts, namely Feri Amsari, Bivitri Susanti and Zainal Arifin Mochtar. This film explains how fraud in elections in Indonesia can damage democracy, especially ahead of the 2024 elections. Acts of fraud create injustice in elections and this is certainly done for personal victory.

The emergence of the dirty vote film received public attention and went viral on social media after being released and watched by millions of people in a short time. The dirty vote film became a topic of conversation, one of which was on the Twitter social media platform. Various comments and responses from the public regarding this film also occurred on Twitter social media. With the diverse sentiments of the public on Twitter social media regarding the dirty vote film, this sentiment analysis research was conducted to determine the views or tendencies of public opinion regarding the dirty vote film which will be classified into positive, negative and neutral sentiments.

Several researchers have used the K-Nearest Neighbor (K-NN) method in sentiment analysis research. Hikmah et al., 2024 recorded a K-NN accuracy of 85.90% with k=6, and a neutral sentiment polarity reaching 86.94% [4]. Palepa et al., 2024 reported an accuracy of 84.65% with a precision of 87% and a recall of 86% [5]. Wijaya & Suwandhi, 2024 showed higher results with an accuracy of 91%, a precision of 93%, and an f-1 score of 92% [6]. The superiority of K-NN in sentiment analysis is found in research by Sandi et al., 2023 showing that K-NN obtained higher results with an accuracy rate of 99% compared to Naïve Bayes which reached 96% [2]. This shows that K-NN has advantages in terms of simplicity and flexibility compared to other methods. By using K-NN, this study seeks to obtain accurate and relevant sentiment analysis results, as well as strengthen existing methodologies in the field of sentiment analysis on social media.

Based on several previous studies, researchers will conduct an analysis using the K-NN algorithm. Researchers also raise a different topic from previous studies, namely the film Dirty Vote, which is a documentary that explains fraud in the 2024 general election in Indonesia. In this study, researchers will analyze public sentiment towards the film Dirty Vote with data obtained from Twitter social media and group opinions into positive, negative, and neutral sentiments. Furthermore, researchers will compare the results of the levels of accuracy, precision, recall, and f-measure from sentiment analysis using the K-NN method. This study aims to understand how the public responds to the film Dirty Vote, which discusses important issues related to fraud in the general election and to apply and evaluate the effectiveness of the K-NN method in sentiment analysis.

2. RESEARCH METHOD

Research methods are a series of procedures and techniques used to collect, analyze, and interpret data to test hypotheses. This method ensures that research is carried out in a systematic, structured manner.

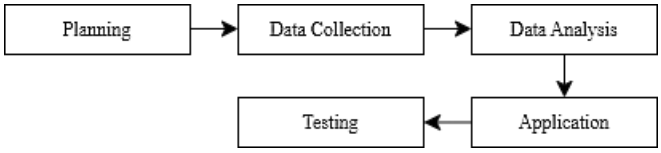

Figure 1. Research Framework

This research begins with the planning stage, where the research topic is determined. The topic discussed in this study is public opinion regarding a documentary film entitled Dirty Vote which was uploaded on social media Twitter. The film raises the issue of fraud in the 2024 general election in Indonesia. After determining the topic, the next step is data collection, which is the process of obtaining the information needed for research. The data collected includes tweets about the Dirty Vote film taken from Twitter. The data collection process was carried out using the Tweet Harvest library in the Python programming language on Google Colaboratory Web. Tweet Harvest is a library used to collect tweet data from social media, especially Twitter. After the data is collected, the next stage is data analysis. Research data analysis involves the process of analyzing, interpreting, and presenting the data that has been collected. In this study, the tweet data obtained

will be grouped into three sentiment classes positive, negative, and neutral. The flowchart system for data analysis in this study can be seen in Figure 2.

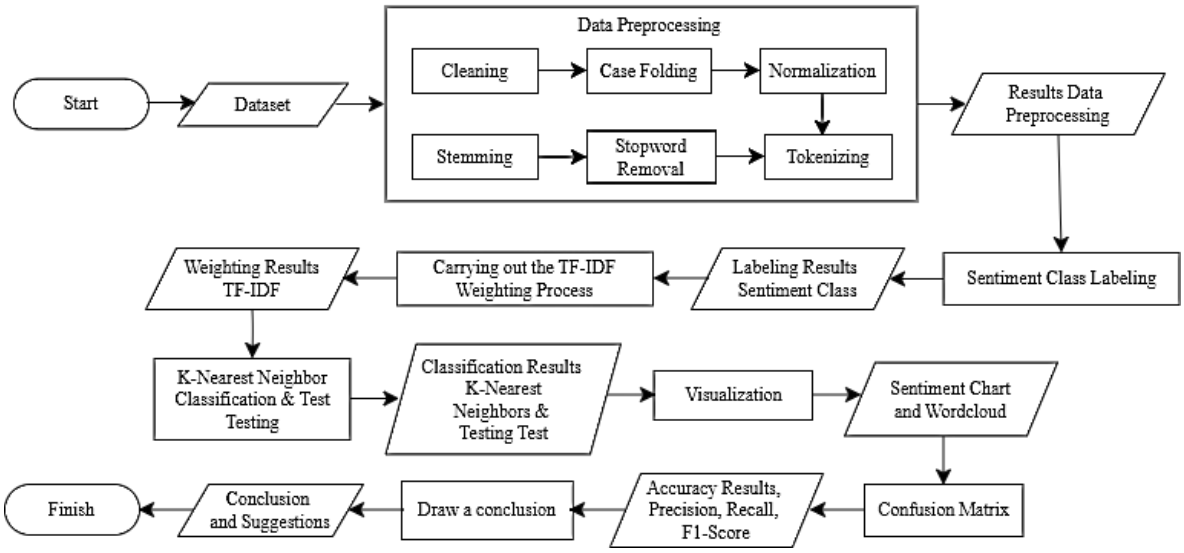

Figure 2. System Flowchart

The application in this study is to find out and analyze the opinions of Twitter users regarding the documentary film entitled dirty vote which explains the alleged fraud in the 2024 election. This study uses the K-Nearest Neighbor (K-NN) method and accuracy level testing. The results of the accuracy test are visualized in the form of a confusion matrix. This research test is to determine the views or sentiments of the Indonesian people towards the dirty vote film on social media Twitter. The testing phase in this study was carried out using the Python programming language in the Google Colaboratory application. The text data analysis process requires a word weighting step, which in this study uses the Term Frequency-Inverse Document Frequency (TF-IDF) method. After the word weighting process, classification is carried out using the K-NN method, and distance calculations are carried out using the Euclidean concept. This approach ensures that sentiment analysis of the Dirty Vote film can be carried out with high accuracy and provides a deep understanding of public reactions on Twitter.

2.1 Text Mining

Text mining is a process of digging up information where users use tools to interact with a number of documents. The aim of this process is to obtain useful information from a number of documents. The concept of text mining is generally applied in separating text documents into categories based on the topics they contain. Text mining can provide solutions to problems such as processing, grouping and analyzing large amounts of unstructured text [7]. Text mining can also be explained as the process of extracting information from data in the form of text or documents. With the aim of finding words that can represent what is in the document so that relationship analysis can be carried out in text mining [8]. The sentiment analysis method requires data collection techniques, in this case the data used is Indonesian text taken from Twitter tweets using a text mining approach and data classification using a machine learning approach [9].

2.2 Sentiment Analysis

Sentiment analysis or often referred to as opinion mining is a combination of several techniques, including Natural Language Processing (NLP), Information Retrieval (IR), and Data Mining (DM) which process or analyze opinions, sentiments and emotions expressed in text form. on an entity. In sentiment analysis, classification models can be used to determine sentiment into two or more classes [10]. This analysis process is usually carried out automatically through several applications or platforms that support text analysis [11]. Sentiment analysis will group the polarity of the text in a sentence or document to determine whether the opinions expressed in the sentence or document are positive, negative or neutral [12]. The sentiment analysis process is influenced by the dataset used. For datasets that consist of a collection of sentences that are quite long, they require different handling [13].

2.3 Preprocessing

Preprocessing is a process where datasets collected from Twitter social media will be cleaned of unnecessary elements so that later we will get data that has quality and matches what the researcher wants [14]. Preprocessing is the selection stage of checking the text by cleaning the text, correcting errors in the text and simplifying the text for further stages. Preprocessing aims to improve the text quality of the data. A model's ability to learn and generalize is affected by data quality [15]. Preprocessing includes cleaning, case folding, normalization, stopword removal, tokenizing, stemming.

2.4 TF-IDF Weighting

In the word weighting stage of the K-NN algorithm, the method used is TF-IDF. TF-IDF weighting is a method of calculating each word that appears in the document data. Document data will be divided into the number of words (terms) in the document data that will be used in classification. Giving weight with TF-IDF aims to analyze how important a word represents a sentence. This weighting is done based on the frequency of occurrence of words in a document [16]. The equations for TF-IDF are as follows.

$$\text{idf} = \log \frac{N}{df} \tag{1}$$

$$w = tf \times idf \tag{2}$$

2.5 K-Nearest Neighbor (K-NN)

K-NN is one of the simplest algorithms for solving classification problems. This algorithm is often used for text and data classification [17]. K-Nearest Neighbor (K-NN) is an algorithm for classifying new objects based on their closest neighbors and the class that appears the most will become the classification result class [18]. Algorithm, which is calculated from the distance value on testing data with training data, from the value of the smallest nearest neighbor [19].

Distance calculations are carried out using Euclidean concepts. The largest number of classes with the closest distance will be the class where the evaluation data is located [20]. The Euclidean distance technique works by finding the shortest distance between two objects without looking at obstacles in the path they follow [21]. The advantage of the K-Nearest Neighbor (K-NN) method is that it is effective on noisy data and effective when the training data is large [22].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

2.6 K-Fold Cross Validation

K-fold cross validation is a process of evaluating the performance of a classification model in a machine learning problem. K-fold cross validation is used to measure the extent to which the model being developed is able to generalize on data that has never been seen before. K-fold cross validation breaks the dataset into several folds and performs several training and testing iterations [23]. K-fold cross validation provides a more consistent picture of system performance, because it involves the entire dataset as part of the testing and training data in several test iterations [24].

In this research, the author divided the previously randomized data into 10-fold models of the same size, so that 9-fold for training data and 1-fold for test data by carrying out 10 iterations. The k-fold cross validation measurement results are the average results of the 10 iterations that have been carried out.

2.7 Confusion Matrix

Confusion matrix is a table that states the classification of the number of correct test data and the number of incorrect test data [25]. Confusion matrix can be used to measure the performance of a classification model and assist in making appropriate decisions to improve its performance.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \tag{4}$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \tag{6}$$

$$\text{F1-Score} = \frac{2 \text{ x precission x recall}}{\text{Precission + recall}} \text{ x } 100\% \quad (7)$$

3. RESULTS AND ANALYSIS

The data taken in this research is the result of crawling data from February 13 2024 to March 11 2024 on Twitter social media using the keyword "film dirty vote" and succeeded in collecting 4000 raw data. The tweet data taken is only in text form and does not contain images. The data collected at the crawling stage is in the form of Indonesian language text taken from tweets from Twitter social media users without any restrictions on the user's age, then the raw tweet data will be saved in Comma Separated Values (CSV) format. In this study, tweet data was classified using the K-Nearest Neighbor (K-NN) method to determine the results of the level of accuracy, precision, recall, f1-score in this study.

Table 1. Sentiment Data

Sentiment
@zainul_munas @DirtyVote Dirty Vote merupakan sebuah film dokumenter yang menggambarkan betapa sudah rusaknya demokrasi yang terjadi di Indonesia. #OrdeBaruBangkit #DirtyVote @aniesbaswedan @cakimiNOW
@tempodotco Buktinya nonton film dokumen dirty vote lalu di validasi ulang jika tidak valid berarti fitnah jika valid langkah terbaik mawas diri karena bersinggungan dgn rakyat Indonesia
@akupadi5 Kemarin katanya film dirty vote bukan kampanye setelah kalah baru ngamuk
Kalau hati sudah mati telinga sudah tuli dan mata sudah buta maka fakta keras di film Dirty Vote pun akan disebut fitnah.
Dirty Vote tanda Indonesia baik-baik saja. Tanda kebebasan berpendapat masih ada. Kalau tidak mungkin tak sempat rekaman apalagi rilis film gratisan.
Masa Tenang Pemilu Tayang Film Dirty Vote tiga Akademisi Dinilai dpt Merusak Demokrasi Indonesia https://t.co/5s8q3YOWOf https://t.co/QcrCyuS0MC

3.1 Preprocessing

After the data has been successfully collected, it will then be processed to the next process stage, namely the data preprocessing process. Preprocessing is the selection stage of checking the text by cleaning the text, correcting errors in the text and simplifying the text for further stages. Preprocessing aims to improve the text quality of the data. Preprocessing includes cleaning, case folding, normalization, tokenizing, stopword removal, stemming.

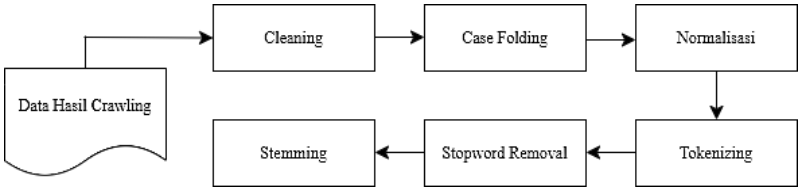

Figure 3. Preprocessing Flowchart

Table 2. Sample Data

Sample Data
@zainul_munas @DirtyVote Dirty Vote merupakan sebuah film dokumenter yang menggambarkan betapa sudah rusaknya demokrasi yang terjadi di Indonesia. #OrdeBaruBangkit #DirtyVote @aniesbaswedan @cakimiNOW

Table 3. Preprocessing Results

Stage	Results
Cleaning	Dirty Vote merupakan sebuah film dokumenter yang menggambarkan betapa sudah rusaknya demokrasi yang terjadi di Indonesia
Case Folding	dirty vote merupakan sebuah film dokumenter yang menggambarkan betapa sudah rusaknya demokrasi yang terjadi di indonesia
Normalization	dirty vote merupakan sebuah film dokumenter yang menggambarkan betapa sudah rusaknya demokrasi yang terjadi di indonesia
Tokenizing	'dirty', 'vote', 'merupakan', 'sebuah', 'film', 'dokumenter', 'yang', 'menggambarkan', 'betapa', 'sudah', 'rusaknya', 'demokrasi', 'yang', 'terjadi', 'di', 'indonesia'
Stopword Removal	'dirty', 'vote', 'film', 'dokumenter', 'menggambarkan', 'betapa', 'rusaknya', 'demokrasi', 'indonesia'
Stemming	dirty vote film dokumenter gambar betapa rusak demokrasi indonesia

3.2 Sentiment Labeling

Sentiment labeling is an important stage in sentiment analysis that aims to identify and classify opinions contained in text. In this research, sentiment labeling was carried out using the text blob method. Sentiment data will be translated into English to get more optimal results. This sentiment labeling is done

automatically and will be given a class label based on its subjectivity and polarity values. This labeling is carried out to determine positive, negative and neutral sentiment classes.

Table 4. Sentiment Labeling

Sentiment	Subjectivity	Polarity	Label
dirty vote documentary film depicting how damaged Indonesian democracy is	0.4	-0.3	Negatif

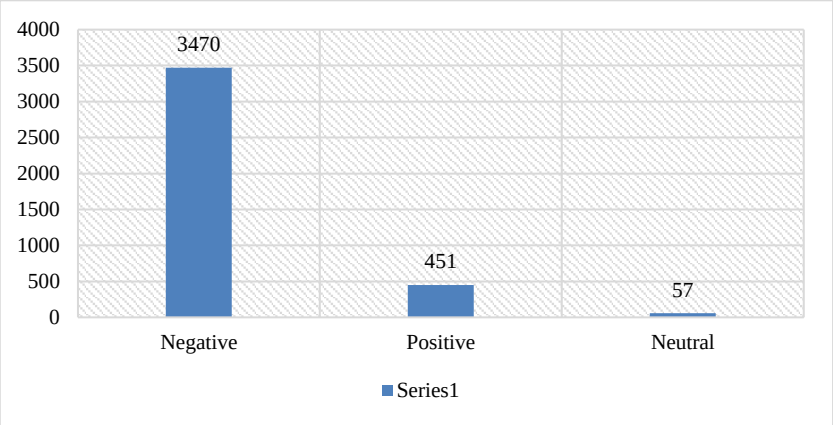

Figure 4. Sentiment Data Classification

Based on Figure 4 above, it can be seen from the sentiment class data for the dirty vote film, which consists of 3978 sentiment data, that the negative sentiment results were 3470 data, positive sentiment was 451 data, and neutral sentiment was 57 data. Therefore, many Twitter users gave negative responses regarding the film Dirty Vote. Below are the results visualization of overall opinion tweet data with negative, positive and neutral sentiment.

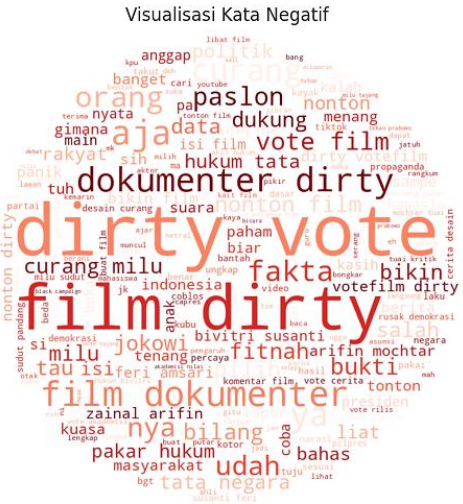

Figure 5. Wordcloud Negative

A word cloud for negative sentiment displays words that frequently appear in text that are considered negative. These words are usually related to complaints, criticism, or feelings of dissatisfaction. The size of a word in a word cloud shows how often the word appears in text that has a negative sentiment. Analysis of these words provides insight into certain issues or aspects of the film that were a source of dissatisfaction or criticism from the public.

A word cloud for positive sentiment displays words that frequently appear in text that are considered positive. These words are usually related to praise, support, or feelings of satisfaction. These words indicate aspects of the film that received appreciation or praise from the audience. This analysis helps identify elements that viewers value or find positive.

The word cloud for neutral sentiment displays words that frequently appear in text that contain no obvious sentiment, either positive or negative. These words may be descriptive or factual and do not express a strong opinion. These words help understand the aspects of the film being discussed objectively, without passing a strong emotional judgment.

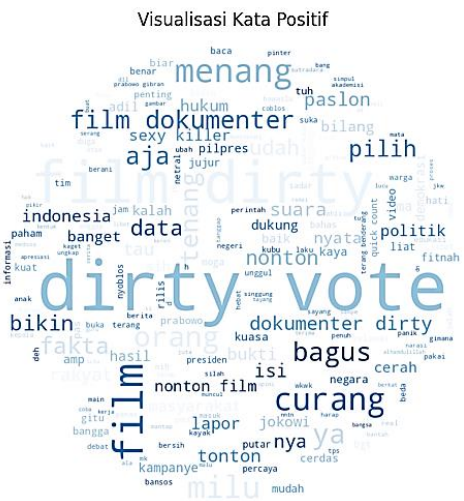

Figure 6. Wordcloud Positive

Figure 7. Wordcloud Neutral

3.3 Word Weighting

Term Frequency-Inverse Document Frequency (TF-IDF) is a method used to measure the importance of a word in a document. In sentiment analysis, TF-IDF helps to understand how relevant a word is in the context of the sentiment analysis being carried out. Inverse Document Frequency (IDF) calculates a term contained in a document. The formula used to calculate IDF is as table 5.

Table 5. Example of the Inverse Document Frequency Calculation Process

Term	TF						DF	IDF
	D1	D2	D3	D4	D5	D6		
dirty	1	1	1	1	1	1	6	0
vote	1	1	1	1	1	1	6	0
film	1	1	1	1	1	1	6	0
dokumenter	1	0	0	0	0	0	1	0.77815125
gambar	1	0	0	0	0	0	1	0.77815125
betapa	1	0	0	0	0	0	1	0.77815125
rusak	1	0	0	0	0	1	2	0.47712125
demokrasi	1	0	0	0	0	1	2	0.47712125
indonesia	1	1	0	0	1	1	4	0.17609125
bukti	0	1	0	0	0	0	1	0.77815125
nonton	0	1	0	0	0	0	1	0.77815125
dokumen	0	1	0	0	0	0	1	0.77815125
validasi	0	1	0	0	0	0	1	0.77815125
ulang	0	1	0	0	0	0	1	0.77815125
valid	0	2	0	0	0	0	1	0.77815125
fitnah	0	1	0	1	0	0	2	0.47712125
langkah	0	1	0	0	0	0	1	0.77815125
baik	0	1	0	0	0	0	1	0.77815125
mawas	0	1	0	0	0	0	1	0.77815125
singgung	0	1	0	0	0	0	1	0.77815125
rakyat	0	1	0	0	0	0	1	0.77815125
kemarin	0	0	1	0	0	0	1	0.77815125
kampanye	0	0	1	0	0	0	1	0.77815125
kalah	0	0	1	0	0	0	1	0.77815125
ngamuk	0	0	1	0	0	0	1	0.77815125
hati	0	0	0	1	0	0	1	0.77815125
mati	0	0	0	1	0	0	1	0.77815125
telinga	0	0	0	1	0	0	1	0.77815125
tuli	0	0	0	1	0	0	1	0.77815125
mata	0	0	0	1	0	0	1	0.77815125
buta	0	0	0	1	0	0	1	0.77815125
fakta	0	0	0	1	0	0	1	0.77815125
keras	0	0	0	1	0	0	1	0.77815125

Term	TF						DF	IDF
	D1	D2	D3	D4	D5	D6		
tanda	0	0	0	0	2	0	1	0.77815125
bebas	0	0	0	0	1	0	1	0.77815125
dapat	0	0	0	0	1	0	1	0.77815125
rekam	0	0	0	0	1	0	1	0.77815125
rilis	0	0	0	0	1	0	1	0.77815125
gratis	0	0	0	0	1	0	1	0.77815125
tenang	0	0	0	0	0	1	1	0.77815125
milu	0	0	0	0	0	1	1	0.77815125
tayang	0	0	0	0	0	1	1	0.77815125
akademisi	0	0	0	0	0	1	1	0.77815125
nilai	0	0	0	0	0	1	1	0.77815125

After determining the Tf and Idf values, then determine the weight of each word in the document. On The weight calculation is carried out by calculating the d-th document against the word (term). Meanwhile, tf is the number of occurrences of term (t) in document (d). IDF is obtained from results of previous process calculations. The results of the term frequency are multiplied by the IDF results will get each weight of each word. The formula used to calculate weight is as follows:

Table 6. Example of Weight Calculation

TF					
D1	D2	D3	D4	D5	D6
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0.77815125	0	0	0	0	0
0.77815125	0	0	0	0	0
0.77815125	0	0	0	0	0
0.47712125	0	0	0	0	0.47712125
0.47712125	0	0	0	0	0.47712125
0.17609125	0.17609125	0	0	0.17609125	0.17609125
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	1.5563025	0	0	0	0
0	0.47712125	0	0.47712125	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0.77815125	0	0	0	0
0	0	0.77815125	0	0	0
0	0	0.77815125	0	0	0
0	0	0.77815125	0	0	0
0	0	0.77815125	0	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0.77815125	0	0
0	0	0	0	1.5563025	0
0	0	0	0	0.77815125	0
0	0	0	0	0.77815125	0
0	0	0	0	0.77815125	0
0	0	0	0	0.77815125	0
0	0	0	0	0.77815125	0
0	0	0	0	0	0.77815125
0	0	0	0	0	0.77815125
0	0	0	0	0	0.77815125
0	0	0	0	0	0.77815125
0	0	0	0	0	0.77815125

3.4 K-Nearest Neighbor Classification

In carrying out the K-Nearest Neighbor classification, first determine the k value, where the k used in this research is k = 4. The calculation is done by calculating the distance between the test data and the training data. After that, sort the distance from smallest to largest value based on the k value, so that the 4 smallest values are taken. This process ensures that classification decisions are based on the most relevant nearest neighbors. as shown in Table 7.

Table 7. Euclidean Distance

Document	Euclidean Distance	Class	Ranking
D1	2.20094410	Negative	1
D3	2.43638499	Negative	2
D4	2.93013549	Negative	3
D5	2.98873895	Positive	4

After sorting it can be seen which classes often appear. Based on Table 7, the neutral class does not appear, the negative class appears 3 times, while the positive class appears 1 time. Based on these results, it can be determined that Document D6 is classified as negative class.

3.5 K-Fold Cross Validation

The 10-fold cross validation test was carried out to determine the validity of the data by testing the model on several different data subsets with 10 iterations. So this test produces 10 output scores for each iteration. Below in Figure 8 and Table 8 present the results of testing the validity of the 10-fold cross validation data.

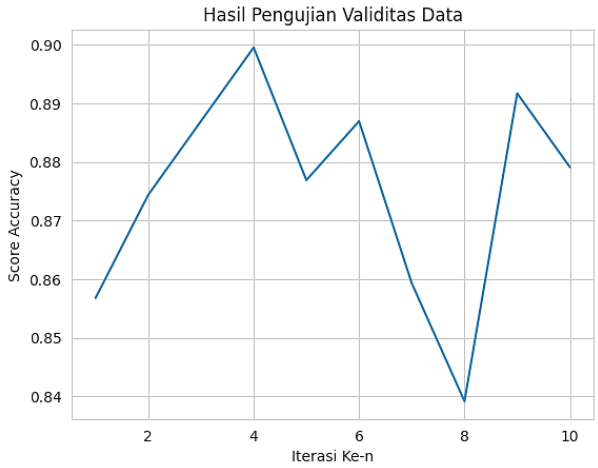

Figure 8. Testing Data Validity

From the graph in Figure 8, it can be seen that the model performance is relatively consistent across all 10 folds, with minimal fluctuation. Each point on the graph represents the results of model evaluation on a particular fold, with the line connecting the points indicating the trend of model performance across folds. This indicates that the model is able to generalize well across multiple data subsets and is not too dependent on a particular data. For example, if the performance line shows little variation between folds, this indicates that the model has high stability and is reliable across multiple conditions.

Table 8. Testing Data Validity

Iteration	Score Accuracy
1	0.85678392
2	0.87437186
3	0.88693467
4	0.89949749
5	0.87688442
6	0.88693467
7	0.85929648
8	0.83919598
9	0.89168766
10	0.8790932
Average	0.87506803

Based on Table 8, the highest accuracy score was in the 4th iteration, namely 0.89949749 89.9%, while the lowest accuracy score was in the 8th iteration, namely 0.83919598 or 83.9%. Overall the average accuracy result in each iteration is 0.87506803 or 87.5%. This indicates that the model has stable and reliable performance across multiple data subsets. Small fluctuations in accuracy values between folds indicate that the model is able to generalize well across multiple data subsets

3.6 Evaluation of Results

In evaluating the results in this study, a confusion matrix was used. In the confusion matrix, accuracy, precision, recall and f1-score values will be calculated to build a classification model using the K-NN method.

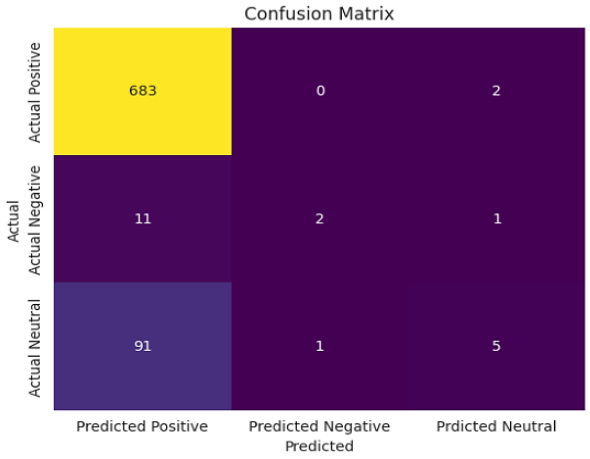

Figure 9. Confusion Matrix

Based on Figure 9, results can be obtained regarding the K-Nearest Neighbor classification test on test data from which accuracy, precision, recall and f1-score values will be calculated using the following equation.

$$\text{Accuracy} = \frac{683+2+5}{683+0+2+11+2+1+91+1+5} \text{ x } 100\% = 87\%$$

$$\text{Precision} = \frac{683}{683+11+91} \text{ x } 100\% = 87\%$$

$$\text{Recall} = \frac{683}{683+0+2} \text{ x } 100\% = 100\%$$

$$\text{F1-score} = \frac{2 \text{ x } 87 \text{ x } 100}{87+100} \text{ x } 100\% = 93\%$$

Based on the results of the confusion matrix calculation, it can be seen that the performance of the K-Nearest Neighbor classification produces an accuracy level of 87%, a precision of 87%, a recall of 100% and an f1-score of 93%. The results of this calculation will be displayed in Figure 10 below:

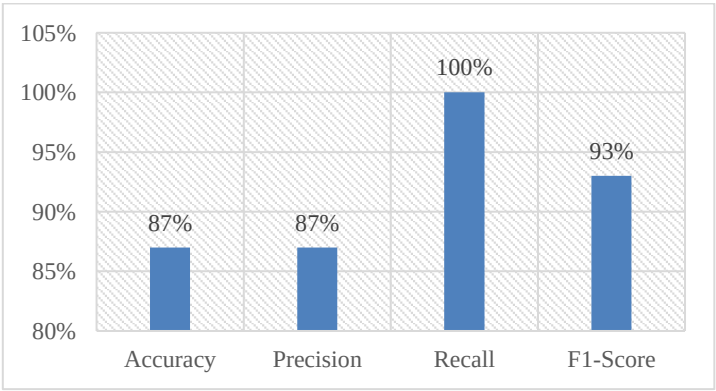

Figure 10. Calculation Results

4. CONCLUSION

The results of sentiment analysis regarding dirty vote films on Twitter social media using the K-Nearest-Neighbor algorithm with the amount of sentiment data collected in the data crawling process amounting to 4000 data. Then after preprocessing there were 3978 data. Labeling carried out using textblob shows that the negative sentiment class is 3470 superior to the positive sentiment class which is 451 and the neutral sentiment class is 57. With these results it can be seen that Twitter social media users give more bad responses in giving opinions on the dirty vote film. The results of the 10-fold cross validation test on the K-Nearest Neighbor algorithm obtained the highest accuracy in the 4th iteration, namely 89.9%, while the lowest accuracy score was in the 8th iteration, namely 83.9%. Overall the average accuracy result in each iteration is 87.5%. Testing in this research was carried out with 80% training data and 20% test data consisting of 3182 training data and 796 test data. In the K-Nearest Neighbor classification which was carried out with k = 4, the accuracy value obtained was 87%, precision was 87%, recall was 100%, and f1-score was 93%.

REFERENCES

[1] B. Huda and B. Priyatna, "Penggunaan Aplikasi Content Management System (CMS) Untuk Pengembangan Bisnis Berbasis E-commerce," *Systematics*, vol. 1, no. 2, p. 81, Dec. 2019, doi: 10.35706/sys.v1i2.2076.

[2] D. Sandi, E. Utami, and K. Kusnawi, "Analisis Sentimen Publik Terhadap Elektabilitas Ganjar Pranowo di Tahun Politik 2024 di Twitter dengan Algoritma KNN dan Naïve Bayes," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, p. 1097, Jul. 2023, doi: 10.30865/mib.v7i3.6298.

[3] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 1, p. 147, Feb. 2021, doi: 10.25126/jtiik.0813944.

[4] E. Maria and E. Junirianto, "Sistem Pendukung Keputusan Pemilihan Bibit Karet Menggunakan Metode Topsis," *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, vol. 16, no. 1, p. 7, Mar. 2021, doi: 10.30872/jim.v16i1.5132.

[5] M. J. Palepa, N. Pratiwi, and R. Q. Rohmansa, "Analisis Sentimen Masyarakat Tentang Pengaruh Politik Identitas Pada Pemilu 2024 Terhadap Toleransi Beragama Menggunakan Metode K - Nearest Neighbor," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 1, pp. 389–401, Feb. 2024, doi: 10.29100/jipi.v9i1.4957.

[6] R. Wijaya and A. Suwandhi, "Sentimen Komentar Universitas Pelita Harapan Pada TikTok Menggunakan Metode K-Nearest Neighbor," *JDMIS: Journal of Data Mining and Information Systems*, vol. 2, no. 1, pp. 26–36, Feb. 2024, doi: 10.54259/jdmis.v2i1.2418.

[7] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter," *Journal of Information System Management (JOISM)*, vol. 3, no. 2, pp. 44–49, Jul. 2021, doi: 10.24076/joism.2021v3i2.558.

[8] Betesda, "Peningkatan Optimasi Sentimen Dalam Pelaksanaan Proses Pemilihan Presiden Berdasarkan Opini Publik dengan Menggunakan Algoritma Naive Bayes dan Paricle Swarm Optimization," *Jurnal Sistem Informasi Universitas Suryadarma*, vol. 7, no. 2, pp. 101–114, Jun. 2014, doi: 10.35968/jsi.v7i2.452.

[9] N. S. Marga, "Sentimen Analisis Tentang Kebijakan Pemerintah Terhadap Kasus Corona Menggunakan Metode Naive Bayes," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 2, no. 4, pp. 453–463, Feb. 2022, doi: 10.33365/jatika.v2i4.1602.

[10] L. A. Andika, P. A. N. Azizah, and R. Respatiwulan, "Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier," *Indonesian Journal of Applied Statistics*, vol. 2, no. 1, p. 34, Jul. 2019, doi: 10.13057/ijas.v2i1.29998.

[11] Herwinsyah and A. Witanti, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *Jurnal Sistem Informasi dan Informatika (Simika)*, vol. 5, no. 1, pp. 59–67, Feb. 2022, doi: 10.47080/simika.v5i1.1411.

[12] A. Z. Amrullah, A. Sofyan Anas, and M. A. J. Hidayat, "Analisis Sentimen Movie Review Menggunakan Naive Bayes Classifier Dengan Seleksi Fitur Chi Square," *Jurnal Bumigora Information Technology*, vol. 2, no. 1, pp. 40–44, 2020, doi: 10.30812/bite.v2i1.804.

[13] W. Widayat, "Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning," *Jurnal Media Informatika Budidarma*, vol. 5, no. 3, p. 1018, Jul. 2021, doi: 10.30865/mib.v5i3.3111.

[14] S. F. Pane and J. Ramdan, "Pemodelan Machine Learning : Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter," *Jurnal Sistem Cerdas*, vol. 5, no. 1, pp. 12–20, May 2022, doi: 10.37396/jsc.v5i1.191.

[15] W. J. Sari *et al.*, "Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 34–43, Apr. 2024, doi: 10.57152/predatecs.v2i1.1119.

[16] S. Suryani, M. F. Fayyad, D. T. Savra, V. Kurniawan, and B. H. Estanto, "Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 1, no. 1, pp. 1–9, 2023, doi: 10.57152/predatecs.v1i1.814.

[17] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, "Analisis Sentimen Aplikasi Ruang Guru di Twitter Menggunakan Algoritma Klasifikasi," *Jurnal Teknoinfo*, vol. 14, no. 2, p. 115, Jul. 2020, doi: 10.33365/jti.v14i2.679.

[18] S. G. Setyorini and Mustakim, "Application of The Nearest Neighbor Algorithm for Classification of Online Taxibike Sentiments In Indonesia In The Google Playstore Application," *Journal of Physics: Conference Series*, vol. 2049, no. 1, p. 012026, Oct. 2021, doi: 10.1088/1742-6596/2049/1/012026.
[19] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 2, no. 1, pp. 26–33, Apr. 2024, doi: 10.57152/predatecs.v2i1.1110.
[20] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74–80, Mar. 2023, doi: 10.33365/jatika.v4i1.2468.
[21] I. Saputra and D. A. Kristiyanti, *Machine Learning Untuk Pemula*. Bandung: Informatika Bandung, 2022.
[22] A. Firdaus, "Aplikasi Algoritma K-Nearest Neighbor pada Analisis Sentimen Omicron Covid-19," *Jurnal Riset Statistika*, pp. 85–92, Dec. 2022, doi: 10.29313/jrs.v2i2.1148.
[23] A. Masruriyah, H. Novita, C. Sukmawati, A. Ramadhan, S. Arif, and B. Dermawan, "Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung," *Computer Science (CO-SCIENCE)*, vol. 4, no. 1, pp. 62–70, Jan. 2024, doi: 10.31294/coscience.v4i1.2389.
[24] D. Nasien, S. Sirvan, D. Deny, R. S. Ryan Syahputra, A. Akbar Marunduri, and R. Prawinata See, "Klasifikasi Penyakit Jantung Menggunakan Decision Tree dan KNN Menggunakan Ektraksi Fitur PCA," *JEKIN - Jurnal Teknik Informatika*, vol. 4, no. 1, pp. 18–24, Feb. 2024, doi: 10.58794/jekin.v4i1.641.
[25] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," *Jurnal Informatika*, vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.

BIBLIOGRAPHY OF AUTHORS

Annisa Fadillah is a student of the Computer Science Study Program at the State Islamic University of North Sumatra, who is currently pursuing a S.Kom degree with a research focus on data mining, especially in data processing using the K-Nearest Neighbors method with the Python programming language.

Sriani earned her S.Kom degree from STMIK Triguna Dharma Medan. Then, she continued her education by earning an M.Kom degree from Universitas Putra Indonesia YPTK Padang. Currently, she serves as a lecturer in computer science at the State Islamic University of North Sumatra. One of her research interests is data mining.