# Sentiment Analysis from Indonesian Twitter Data Using Support Vector Machine And Query Expansion Ranking

**Hasbi Atsqalani[1], Nur Hayatin[2], Cristian Sri Kusuma Aditya[3]**
[1,2,3]Department of Informatics, University of Muhammadiyah Malang, Indonesia

**Article Info**

***Article history:***

Received February 22, 2021
Revised February 25, 2022
Accepted April 19, 2022
Published June 30, 2022

***Keywords:***

Sentiment analysis
Support vector machine
Query expansion ranking
Social media data

**ABSTRACT**

Sentiment analysis is a computational study of a sentiment opinion and an overflow of feelings expressed in textual form. Twitter has become a popular social network among Indonesians. As a public figure running for president of Indonesia, public opinion is very important to see and consider the popularity of a presidential candidate. Media has become one of the important tools used to increase electability. However, it is not easy to analyze sentiments from tweets on Twitter apps, because it contains unstructured text, especially Indonesian text. The purpose of this research is to classify Indonesian twitter data into positive and negative sentiments polarity using Support Vector Machine and Query Expansion Ranking so that the information contained therein can be extracted and from the observed data can provide useful information for those in need. Several stages in the research include Crawling Data, Data Preprocessing, Term Frequency – Inverse Document Frequency (TF-IDF), Feature Selection Query Expansion Ranking, and data classification using the Support Vector Machine (SVM) method. To find out the performance of this classification process, it will be entered into a configuration matrix. By using a discussion matrix, the results show that calcification using the proposed reached accuracy and F-measure score in 77% and 68% respectively.

***Corresponding Author:***

Nur Hayatin
Informatics Department, Engineering Faculty
University of Muhammadiyah Malang
Jl. Raya Tlogomas 246 Malang, Indonesia
noorhayatin@umm.ac.id

## 1. INTRODUCTION

Indonesia is a country that adheres to a democratic system. This is marked by the general election (election) of presidential and vice presidential candidates. This general election is usually held simultaneously throughout Indonesia every 4 years. In 2019 this is the year when general elections will be held. For the 2019-2024 period there are two presidential and vice presidential candidates who have passed the requirements to run, namely Ir. H. Joko Widodo - Prof. Dr. KH Ma'ruf Amin and Lieutenant General (Purn) H. Prabowo Subianto - H. Sandiaga Uno. The excitement and enthusiasm of the people towards the presidential candidates (CAPRES) and vice presidential candidates (CAWAPRES) is not only visible in the real world, but also in the world of social networking like twitter.

The development of using Twitter is very fast. According to the news portal websindo.com in January 2019, social media users in Indonesia reached 150 million users and Twitter users reached 6.43 million users[1]. Tweets are user status which is used to provide information. The contents of the tweet can be used to express a person's feelings towards the presidential and cawapres candidate pairs, for example "I am very happy with his leadership" this is a subjective assessment or opinion. The opinion in this tweet is used to see how the public's sentiment is for the presidential and vice presidential candidates.

There are various classification techniques to determine public sentiment, including the Naive Bayes classifier, Support Vector Machine, and Decision Trees. In another study that conducted sentiment analysis on Twitter regarding the use of land public transportation in cities using the Support Vector Machine method, the results of these tests obtained an accuracy of 78.12%. [2] In another study on the analysis of the 2013 curriculum sentiment on social media twitter using the K-NN method and the Feature Selection Query Expansion Ranking they got an accuracy of 96.36%. [3] In other research on the classification of web pages using the weighted voting support vector machine method, the results obtained an accuracy of 74%. Another research for classification without using feature selection, while in the research conducted by Nurul et al they applied the Query Expansion Ranking feature selection to the K-NN method, the K-NN method had a disadvantage, namely it could not handle the missing value [4]. So that research will be carried out that can cover the shortcomings of the K-NN method and improve the accuracy of the SVM method. The SVM method itself has the advantage of being able to calculate patterns that are not included in a class or category.

## 2. METHOD

The stages in this research are data crawling, text preprocessing, weighting using TF-IDF, Feature selection using Query Expansion Ranking, and finally classifying using Support Vector Machine method. The research flow can be seen in Figure 1.

The stages that will be carried out in this research are as follows:

### 1.1. Data collection

Data to be used in this study is the tweet data of the Indonesian people on the topic of general elections in Indonesia in 2019. The tweet data taken is the one using the hashtag "# 2019tetapjokowi" and "# 2019gantipresiden". The data crawling process was carried out on April 8-17-2019. Data was collected using a crawling process through twitter scraper then it will label manually into two classes positive and negative labels. Total data used in this research is 1500 tweets with balance proportion 750 data for each class: positive and negative.

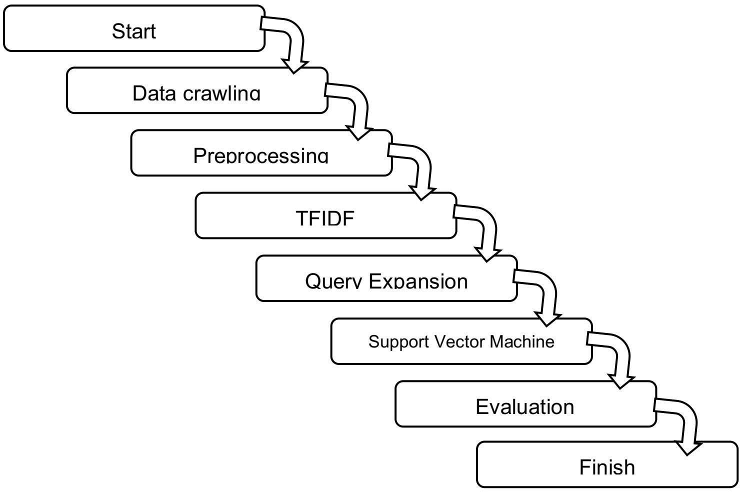

Figure 1. Research Flow

### 1.2. Preprocessing Stage

There are 5 preprocessing stages [5], namely: *Case Folding* i.e. converting all letters in the text to lowercase. In the next stage, punctuation marks and numbers that are not important in the sentiment analysis process will be removed, such as commas (,), periods (.), 1, 2, and others. Tokenizing is a process for sorting the contents of the text into words. Stopword removal is done to remove words that are not used in the sentiment analysis process. Such as "and", "or", and others. Stemming, namely the process of returning the affix to the root word by eliminating prefix and suffix [6]. Stage of the preprocessing process and the results can be seen in Table 1.

Table 1. Stage of the preprocessing process

| Stage | Tweet |
|---|---|
| Original Tweet | Yuk dukung jokowi tahun depan, indonesia damai selalu #2019tetapjokowi |
| Case Folding | yuk dukung jokowi tahun depan indonesia damai selalu #2019tetapjokowi |
| Punctual Removal | yuk dukung jokowi tahun depan indonesia damai selalu |
| Tokenizing | [yuk] [dukung] [jokowi] [tahun] [depan] [indonesia] [damai] [selalu] |
| Stopword Removal | [dukung] [jokowi] [tahun] [depan] [indonesia] [damai] [selalu] |
| Stemming | [dukung] [jokowi] [tahun] [depan] [indonesia] [damai] [selalu] |

### 1.3. Weighting use TF- IDF

After the data passes the preprocessing stage, the data will then go through the word weighting stage using Term Frequency - Inverse Document Frequency (TF-IDF). TF-IDF can be demonstrated in 3 ways [6], using equation (1) to (3).

Using binary values, which is given a value of 1 for words contained in the document and a value of 0 for words that do not appear in the document. In this concept, the frequency of occurrence of words is not included in the calculation, Using the frequency value of the occurrence of words directly to become TF, Using the fractional term values that have been normalized using equation (1):

$$TF(t,d) = 0{,}5 + 0{,}5 \frac{f(t,d)}{\max\{f(w,d):w \epsilon d]}TF(t,d) = 0{,}5 + 0{,}5 \frac{f(t,d)}{\max\{f(w,d):w \epsilon d]} \quad (1)$$

Where the frequency the word appears on is the maximum frequency of other terms in the document. $(t,d)t \max\{f(w,d):w \epsilon d\}d(t,d)t \max\{f(w,d):w \epsilon d\}d$ Meanwhile, to calculate the IDF, the following formula can be presented in equation (2):

$$IDF(t,d) = log \frac{N}{Df(t,d)}IDF(t,d) = log \frac{N}{Df(t,d)} \quad (2)$$

Where N is the number of documents and $Df(t,d)Df(t,d$as many documents in document D containing term t. However, if the term does not appear, there will be a value of 0 in the division, so it needs to be handled to change it to $1 + Df(t,d).1 + Df(t,d).$Calculating TF-IDF from words can be conducted using equation (3):

$$TF - IDF(t,d,D) = tf(t,d)x\ idf(t,D)TF - IDF(t,d,D) = tf(t,d)x\ idf(t,D)$$

(3)

From the formula above, a value will be obtained that can be used as a weighting of words when grouping words.

### 1.4. Expansion Ranking Query Selection Feature

Feature *Selection* is one of the most important factors that can affect the level of classification accuracy. Feature selection is an optimization process to reduce a large set of original source features to a relatively small feature subset that is significant for fast and effective classification accuracy.

*Feature selection* used in this study is the Query Expansion Ranking which is a suggestion from [7]. The Query Expansion Ranking method is based on the Query Expansion technique and the probabilistic weighting model used to score a word. The following equation (4), (5) and (6) respectively show the calculation process used for feature selection.

$$pf = \frac{df_+^f + 0.5}{n^+ + 1.0}pf = \frac{df_+^f + 0.5}{n^+ + 1.0} \quad (4)(4)$$

The variable $pfpf$ is a probability value *term f* on the training data document in the positive category, and $df_+^f$ $df_+^f$ is a number of documents containing *term f* existing in the positive category training data. Meanwhile, $n^+n^+$ is the total number of positive category training data documents .

Furthermore, calculation *qf* using variable $df_{-}^{f} df_{-}^{f}$, a number of documents containing *term f* which is in the negative category training data, and $n^{-} n^{-}$ is the total number of negative category training data documents . Meanwhile $score_f score_f$ is the result of the calculation of the Query Expansion.

$$qf = \frac{qf_{-}^{f}+0.5}{n^{-}+1.0} qf = \frac{qf_{-}^{f}+0.5}{n^{-}+1.0} \qquad (5)(5)$$

$$score_f = \frac{|pf+qf|}{|pf-qf|} score_f = \frac{|pf+qf|}{|pf-qf|} \qquad (6)(6)$$

### 1.5. SVM Classification (Support Vector Machine)

Basically, the SVM algorithm is used for the classification process between two classes or binary classification. Along with its development, SVM is also used for multi-class classification by combining several binary classifiers. SVM is used to find the best hyperplane by maximizing the distance between classes. Hyperplane is a function that can be used to separate between classes. In 2-the functions used for classifications between classes are called rows whereas, the functions used for classifying between classes in 3-D are called similar fields, while the functions used for classifications in higher dimensional classrooms are called hyperplanes. For classification using SVM we follow the research from [8].

### 1.6. Evaluation

After classification using SVM method, intrinsic testing will be carried out. Intrinsic testing is done by calculating the accuracy and f-measure obtained from precision and recall calculations [9]. The equation of accuracy, precision, recall and F-measure respectively is presented in (7), (8), (9), (10).

$$Accuracy = \frac{correctly\ classified\ positive\ tweets\ (TP) + correctly\ classified\ negative\ tweets\ (TN)}{total\ tweets\ (TP + FP + FN + TN)} \qquad (7)$$

$$Precission = \frac{correctly\ classified\ positive\ tweets\ (TP)}{correctly\ classified\ positive\ tweets\ (TP) +\ incorrectly\ classified\ positive\ tweets\ (FP)} \qquad (8)$$

$$Recall = \frac{correctly\ classified\ positive\ tweets\ (TP)}{correctly\ classified\ positive\ tweets\ (TP) +\ incorrectly\ classified\ negative\ tweets\ (FN)} \qquad (9)$$

$$F - Measure = \frac{2 \times Precission \times Recall}{Precission + Recall} \qquad (10)$$

## 3. RESULTS AND DISCUSSION

Total data used in this research is 1500 tweets with balanced proportion 750 data for each class: positive and negative, where the total of train and test data proportion follows each scenario using random technique. This testing process is carried out in 5 test scenarios to get better performance of the model proposed, the support vector machine with query expansion ranking (SVM-QER). The first scenario is testing using SVM-QER with the sharing of test data by 40% and training data by 60%. In the second scenario, there is a test using the support vector machine classification method and the query expansion ranking by dividing the test data by 40% and training data by 60%. The third scenario is the distribution of test data by 50% and training data by 50%. The fourth scenario is the distribution of test data by 60% and training data by 40%. And the fifth testing scenario with the distribution of test data by 80% and training data by 20%. Meanwhile, the confusion matrix table for each scenario is presented in Table 1.

Table 2. Confusion matrix for each Scenario

True Label

| | | Scenario 1 | | Scenario 2 | | Scenario 3 | | Scenario 4 | | Scenario 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | + | - | + | - | + | - | + | - | + | - |
| Predicted | + | 333 | 1 | 12 | 75 | 9 | 118 | 12 | 142 | 9 | 147 |
| Label | - | 193 | 73 | 0 | 193 | 1 | 322 | 1 | 385 | 0 | 484 |

Table 2 compares the evaluation result for each scenario. The results obtained from the best accuracy value reached approximately 77% in the fifth testing scenario with a comparison of 80% for test data and 20% for training data. Then followed by the second and third scenarios with the same accuracy value of 74%. Meanwhile, the first and the second secondary present accuracy values around 61% and 73% respectively. Similar result is shown for the F-measure evaluation result, the fifth scenario produces a higher score in 68%, and the lowest f-measure score is in the first scenario. The ratio of TN values that tend to be dominant or more due to the use of imbalanced datasets, more for tweets with negative sentiments. For more clearly, comparing of evaluation result depicts in Figure 2

Table 3. Evaluation result for all scenario SVM-QER

| | Accuracy | Precision | Recall | F measure |
|---|---|---|---|---|
| Scenario 1 | 61 | 46 | 61 | 53 |
| Scenario 2 | 73 | 81 | 73 | 65 |
| Scenario 3 | 74 | 78 | 74 | 64 |
| Scenario 4 | 74 | 79 | 74 | 64 |
| Scenario 5 | 77 | 82 | 77 | 68 |

Furthermore, we compare the performance between SVM and SVM-QER through analyzing the evaluation result of accuracy, precision, recall, and F-measure for each method. The results of the comparison can be seen in the figure 3. From the bar chart, we can see that the accuracy value and F-measure score of SVM are 61% and 53% respectively. Meanwhile, for SVM-QER the best scenario used with an accuracy, precision, recall, and F-measure are 77%, 82%, 77%, and 68% respectively. Therefore, we can conclude that better accuracy results are obtained when the support vector machine method is combined with the feature selection query expansion ranking with the accuracy value increasing up to 16% while F-measure score increased by 15%. For future work, we will improve the model proposed with additional data and testing with some various preprocessing combination steps.

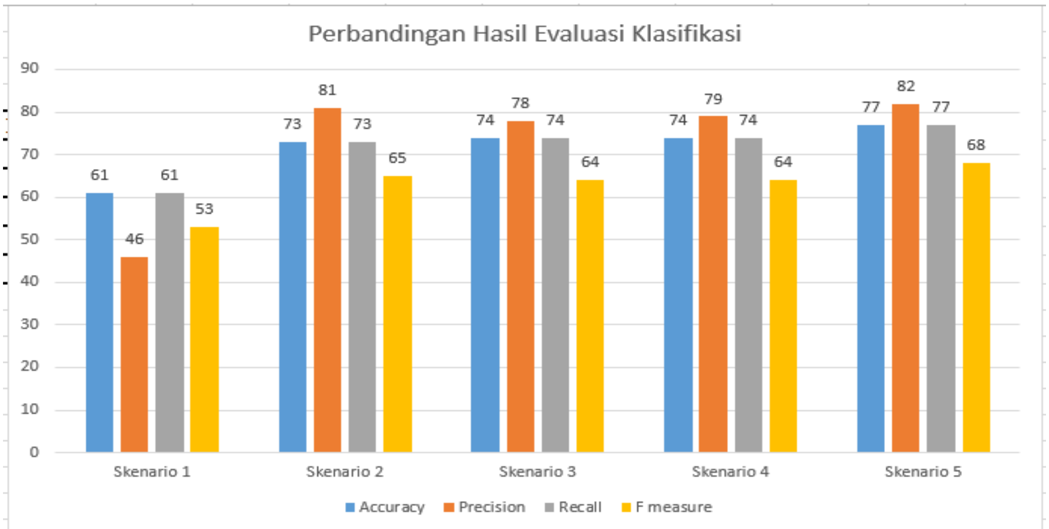

Figure 2. The Bar chart of evaluation result for each scenario

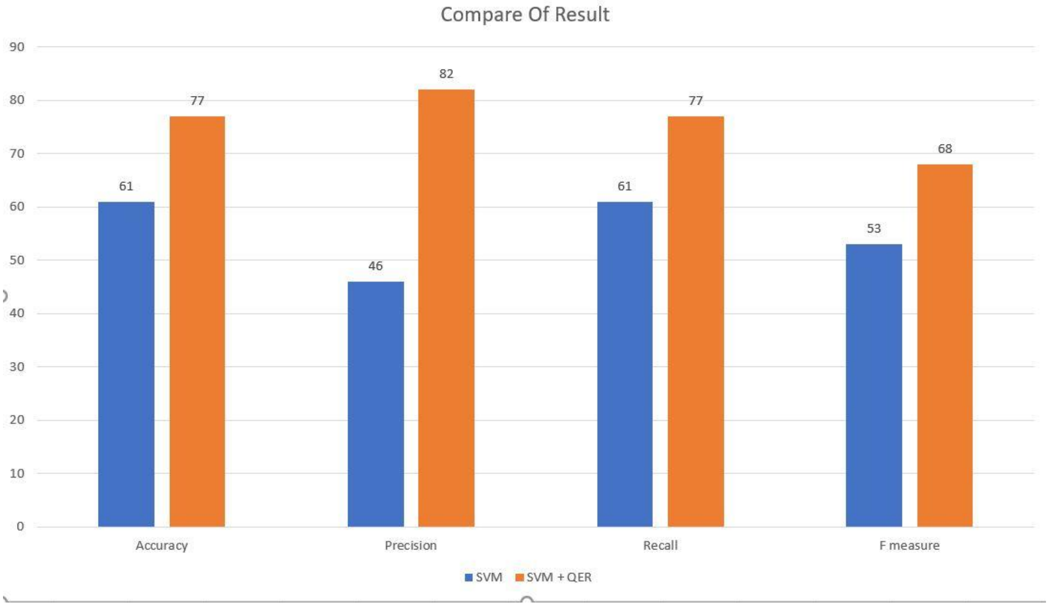

Figure 3. Comparison Results SVM and SVM+QER

## 4. CONCLUSION

The research proposed sentiment analysis model for Indonesian twitter data using combination Support Vector Machine and Query Expansion Ranking. To measure the performance of the model proposed, we compare SVM-QER with SVM and analyze the result of accuracy, precision, recall, and F-measure score for each method. Based on the evaluation, it can be concluded that better accuracy results are obtained when the support vector machine method is combined with the feature selection query expansion ranking with the accuracy value increasing up to 16% while F-measure score increased by 15%.

## ACKNOWLEDGEMENTS

Thanks to those who have helped complete this research. To the Muhammadiyah University of Malang which has provided the facilities to complete this research to achieve what is the goal of this research.

## 5. REFERENCES

[1] O. C. WEBSINDO, “INDONESIA DIGITAL 2019 MEDIA SOSIAL,” 2019. .

[2] R. C. Chen and C. H. Hsieh, "Web page classification based on a support vector machine using a weighted vote schema," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 427–435, 2006, doi: 10.1016/j.eswa.2005.09.079.

[3] N. D. Mentari, M. A. Fauzi, and L. Muflikhah, "Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 8, pp. 2739–2743, 2018.

[4] Anwar, M. Y. (2019). *Klasifikasi Halaman Web Untuk Anak Menggunakan Metode Weighted Voting Support Vector Machine*.

[5] R. Adhitia and A. Purwarianti, "Penilaian Esai Jawaban Bahasa Indonesia Menggunakan Metode Svm - Lsa Dengan Fitur Generik," J. Sist. Inf., vol. 5, no. 1, p. 33, 2012.

[6] Arifin, A. Z., Mahendra, I. P. A. K., & Ciptaningtyas, H. T. (2009). Enhanced confix stripping stemmer and ants algorithm for classifying news document in indonesian language. In *The International Conference on Information & Communication Technology and Systems* (Vol. 5, pp. 149-158).

[7] H. Himawan, W. Kaswidjanti, A. Sentimen, M. Sosial, and L. Based, "Metode Lexicon Based Dan Support Vector Machine Untuk Menganalisis Sentimen Pada Media Sosial Sebagai Rekomendasi Oleh-Oleh Favorit," vol. 2018, no. November, pp. 235–244, 2018.

[8] T. Parlar and S. A. Ozel, "A new feature selection method for sentiment analysis of Turkish reviews," *Proc. 2016 Int. Symp. Innov. Intell. Syst. Appl. INISTA 2016*, no. December 2017, 2016, doi: 10.1109/INISTA.2016.7571833.

[9] H. P. P. R. Zuriel and A. Fahrurozi, "Implementasi Algoritma Klasifikasi Support Vector Machine Untuk Analisa Sentimen Pengguna Twitter Terhadap Kebijakan Psbb," *J. Ilm. Inform. Komput.*, vol. 26, no. 2, pp. 149–162, 2021.

[10] R. Adhitia and A. Purwarianti, "Penilaian Esai Jawaban Bahasa Indonesia Menggunakan Metode Svm - Lsa Dengan Fitur Generik," *J. Sist. Inf.*, vol. 5, no. 1, p. 33, 2012, doi: 10.21609/jsi.v5i1.260.