JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.5 No.2 May 2023: 79-83
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v5i2.9978

Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech

Kristien Margi Suryaningrum
Software Engineering Program, Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
kristien.s@binus.edu

Correspondence: kristien.s@binus.edu

Abstract *– Hate speech is a form of expression used to spread hatred and commit acts of violence and discrimination against a person or group of people for various reasons. Cases of hate speech are very common in social media, one of which is Twitter. The goal to be achieved is to create a system that can classify a tweet on Twitter into hate speech (HS) or non-hate speech (NONHS) classes. The method used is Support Vector Machine by comparing the features of TF-IDF and Count Vectorizer. And the parameters compared are seen from accuracy, precision, recall, and f1-score. Results obtained, overall, by using the TF-IDF feature, the Support Vector Machine algorithm gets high results compared to the Count Vectorizer feature, with an Accuracy Value of 88.77%, 87.45% Precision, 88.77% Recall, and F1-score of 87.81%.*

Keywords: *TF-IDF; Count Vectorizer; Support Vector Machine; Sentiment Analysis*

I. INTRODUCTION

Social media like Twitter is a viral social media in the cyber world. Twitter was founded in 2006 but as of 18 May 2013 (www.statisticbrain.com), it has reached 554,750,000 users, 22 times more than MySpace, another social media founded three years earlier than Twitter. Twitter has its own traits and characteristics, which may be simpler than other social media tools. Many terms exist only on this bird symbol site. First, Twitter operates a digital information service that enables users to send and track micro-messages known as (tweets) of no more than 140 characters. The tweet facility is designed for use on mobile devices and PCs. Twitter is used to post any status, repost other users' statuses (retweets), respond to other users' posts (replays) and share links (Ivan et al, 2019), (Zhang & Luo, 2019).

Bloggers may also have links to useful blog posts sent via Twitter, either post-by-post or automatically, as is the case with all new media, old media is not obliterated but classed and promoted on Twitter. To update the information held by any user or account owner, people can follow (subscribe to) that person's Twitter account. When a Twitter user has followed someone's account, any broadcast information (posts) will appear in an updated feed which is called the timeline. The number of followers (total number of people following) and followers (total number of people following) are always updated in a specific box on the Twitter profile page.

Hate speech is a form of expression used to spread hatred and commit acts of violence and discrimination against a person or group of people for various reasons. Cases of hate speech are very often found on social media, one of which is on Twitter (Modi, 2018).

In classifying hate speech on Twitter, a system is also needed that can help users find out which are hate speeches and which are non-hate speeches by using the right algorithm and feature extraction that will predict the most accurate value. TF-IDF and Count Vectorizer are features to find out how often a word appears in a document. However, using one of the two methods can be measured (based on parameters) which one is the fastest for classifying hate speech using the Support Vector Machine algorithm.

So based on the background of these problems, the main problem that will be raised and studied in this study is how the comparison of TF-IDF and Count Vectorizer features is measured in terms of Accuracy, Precision, Recall, and

F1-score in classifying hate speech in Indonesian on social media Twitter uses the Support Vector Machine Algorithm.

II. METHODS

Based on Figure 1 it can be seen the stages that will later be carried out in this study. These stages are in the form of data collection, data labeling, text preprocessing, feature extraction, model classification, data validation, and data evaluation, and are shown in Figure 1 (Modi, 2018).

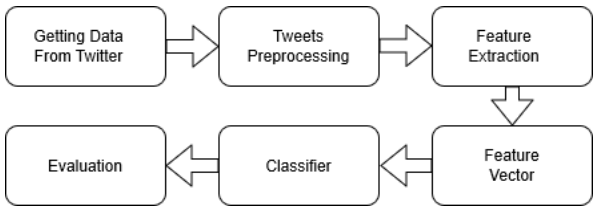


Figure1. Sentiment Analysis Flowchart (Modi, 2018)

2.1. Data Collection

At the data collection stage, using social media Twitter as a medium for collecting the required datasets related to research. The data collected is in the form of Indonesian which is hate speech, in CSV format, from 2020-2022

2.2. Labeling Data

At this stage, the data that has been retrieved through the crawling process will be labeled hate speech and not hate speech. Labeling will be based on predetermined rules. Tweet data must contain words that are included in the hate speech category and the intended subject of the said word also tweet data must have a context that meets the requirements as stipulated in Law No. 19 of 2016 concerning Information and Electronic Transactions Article 28 Paragraph (2).

If a tweet meets the rules that have been set before, then the tweet will be labeled hate speech (1). If a tweet does not meet the rules that have been set before, then the tweet will be labeled non-hate speech (0). This data in the future will be referred to as the tweet dataset (Amri, 2020).

2.3. Text Pre-processing

Text pre-processing is a process in which text will be reduced and processed with the aim of removing words or letters that do not have a contribution that will affect the next process (Amri, 2020). So that the resulting text will be ready to use for shopping sentiment analysis. The stages in text pre-processing are data cleansing, case folding, tokenization, stemming & lemmatization, normalization, and stopword removal.

2.4. Extraction Features

After the data has passed the text pre-processing stage, the next stage is feature extraction. This stage aims to obtain important features to simplify the process of classifying data. The data to be extracted is data that has passed the text pre-processing stage. Furthermore, the extracted dataset will be used as training data for model classification. At this stage, the writer will use two methods, namely Count Vectorizer and Term Frequency – Inverse Document Frequency (TF-IDF).

2.5. Classification Models

After getting the data that has been previously processed, at this stage sentiment classification will be carried out to share the words contained in the tweet in order to obtain the appropriate recognition class or category (Istaiteh et al, 2020).

2.6. Data Validation

The data validation stage used in this study is the K-Fold Cross Validation. This stage is useful for testing security and can shorten processing time, by using subset data. This will be done until all parts of the data subset have been used as tests.

2.7. Data Evaluation

This study uses two prediction classes, namely hate speech and non-hate speech classes. The Confusion Matrix table is used to make it easier to evaluate data. The confusion matrix is a summary table of the number of correct and incorrect predictions. In the confusion matrix, true positives (TP) and true negatives (TN) are used to indicate that the data has been classified correctly according to its class. Meanwhile, false positives (FP) and false negatives (FN) are used to inform that there is data that is classified incorrectly, and are shown in Figure 2 (Moh et al, 2020).

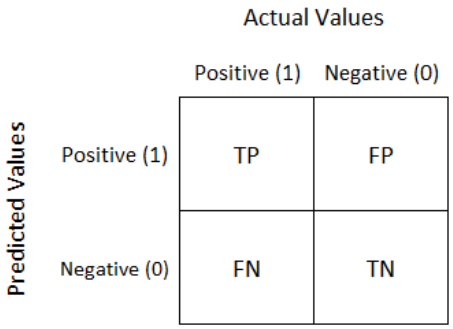


Figure 2. Confusion Matrix

The parameters used to measure are accuracy, memory, precision, and f1 score. Accuracy is a way to calculate the accuracy of a classification result with an actual value. The recall is a way of calculating the true positive rate by calculating the comparison of the number of positive data that are classified as positive with the total positive data events obtained (Watanabe, 2018). Precision is the stage where calculations are carried out to compare the number of positive data that are classified as positive with the total number of data events that are predicted to be positive. Calculation of the f1-score is used to calculate the average comparison of recall and precision.

III. RESULTS AND DISCUSSION

This research process was carried out in several stages. Starting from the process of crawling Twitter data, labeling, text pre-processing, data vectorization using a count vectorizer and TF-IDF, and classification.

3.1. Data Collection

The sampling data used is from Twitter data, from 2020 to 2022. Data taken through a crawling process is data related to hate speech and the Indonesian language. When

doing data crawling the author uses several search keywords which can make it easier to collect hate speech data.

Examples of keywords used when crawling data are "anjing, babi, banci, bangsat, bangsad, bejat, bodo, brengsek, goblok, gembel, gila, keparat, kafir, kampret, monyet, sipit" etc.

The data crawling process is carried out using the Python library which will be used to interact with the existing API on Twitter. The use of this library can also be done to retrieve all data that has been crawled into CSV format so that it makes it easier to process the data that has been retrieved. The resulting tweet data is dirty data that contains data that is less relevant, such as duplicate data. Then, the data was cleaned manually, looking for data that had similar words, so that after cleaning the data, we finally got 100,592 tweets

3.2. Data Labeling

After the first stage, data is collected and data is cleaned. From the results of the crawling process, data is labeled according to its category. The data is categorized into two categories, namely tweets that contain elements of hate speech will be labeled 1, and data that is more towards non-hate speech will be labeled 0.

3.3. Data pre-processing

The third stage is the data pre-processing stage. After the data is finished in the labeling process, then the data will enter the text pre-processing stage. At this stage, it is useful to make the data ready to be processed so that it will not affect the results of the analysis (Adeva et al, 2014), (Ritonga & Purwaningsih, 2018).

The researcher retrieved data in CSV format which had been integrated in the previous stage into Python through a library. This stage aims to convert the data contained in the CSV into an object so that it can be used for the following stages.

At the case folding stage, it is useful to convert data into the same format, namely into all lowercase letters. This conversion is done because the data taken is not uniform, there are several capital letters or there are abbreviations that use capital letters. All capital letters will be lowercase so that they are uniform (Tineges et al, 2020).

The tweet data obtained is dirty data that still contains links, mentions, hashtags, symbols, and punctuation marks. Mention is a term on Twitter where users make calls to other users, mentions begin with the '@' symbol and continue with the username. Then hashtags are used by Twitter users to make indicators of certain topics, hashtags starting with the '#' symbol and followed by a certain word. Symbols that represent emoticons and punctuation such as quotation marks, question marks, and exclamation marks. All kinds of things that have been mentioned above will be omitted in the text data because they are irrelevant and not needed in the classification process (Salsabila et al, 2018).

The tweet data obtained is dirty data that still contains links, mentions, hashtags, symbols, and punctuation marks. Mention is a term on Twitter where users make calls to other users, mentions begin with the '@' symbol and continue with the username. Then hashtags are used by Twitter users to make indicators of certain topics, hashtags starting with the '#' symbol and followed by a certain word. Symbols that represent emoticons and punctuation such as quotation marks, question marks, and exclamation marks. All kinds of things that have been mentioned above will be omitted in the text data because they are irrelevant and not needed in the classification process.

The next stage is the tokenization stage. The purpose of tokenization is to make it easier to process data during the normalization process. The data that has been processed before will be separated per word into an array (Rachmah & Baharuddin, 2019).

This normalization stage will change words that were previously non-standard words into standard words according to the Big Indonesian Dictionary (KBBI), with the aim that at the next stage of the process the results will be more valid (Mustafa et al, 2018).

The next stage is stemming & lemmatization. This stage will transform words into basic words by removing initial and final affixes to words, to help make it easier to remove affixes contained in the text data (Istaiteh et al, 2020). For example, there are many affixes "me", "di", "ke", "per", etc.

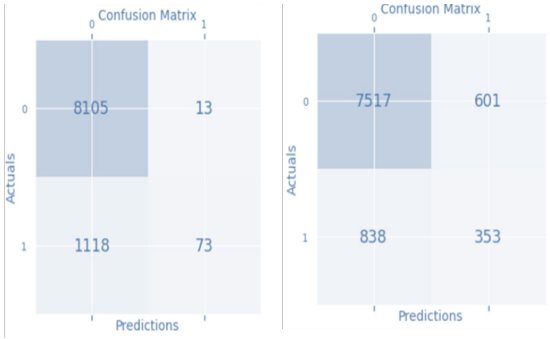
This stage will remove conjunctions or words that are not needed. Researchers use the literary library. This library helps to remove words that can interfere with the process being carried out. This library already provides default stopwords in Indonesian. For example "aku", "kamu", "dari", "kepada", etc.

3.4. Sentiment Analysis Modeling

The 4th stage is Sentiment analysis modeling. In the sentiment analysis modeling stage, data that has been labeled will be categorized into positive and negative sentiments. In the first feature extraction stage, namely the count vectorizer, it is performed to change what was previously a text feature, into a vector representation. The next feature extraction stage is TF-IDF. Through this TF-IDF process, each data will be decomposed into a numeric or vector format. This process makes it easier to read and gives weight to each word so that the results are more accurate. This stage also calculates the Term Frequency and Inverse Document for each word that appears in the data set, using the Tfidf-Vectorizer to calculate calculations related to TF-IDF values.

The K-Fold Cross Validation stage will be used to divide the data samples randomly and group data with as many as k-fold values. This stage of the scoring function is useful for use in the sentiment analysis stage. At this stage, it will display the confusion matrix and the results of the accuracy score, precision score, recall score, and f1-score. In addition, the macro average and weighted average are also displayed (Mustafa et al, 2018).

The stages of the classification model are useful for modeling the Support Vector Machine algorithm. Later the model formed will be used for sentiment analysis, and are shown in Figure 3.



(a) (b)

Figure 3. Sentiment Analysis Modelling using TF-IDF Features (a) and using Count Vectorizer Features (b)

3.5. Evaluation

The research results from the data analysis process that has been carried out show a comparison of the predicted values of the Support Vector Machine (SVM) algorithm, using the Count Vectorizer feature and the TF-IDF feature, and are shown in Table I.

Table I. Comparison of Accuracy, Precision, Recall, F1-score Support Vector Machine Algorithms, Using Count Vectorizer Features and TF-IDF Features

Feature	Accuracy	Precision	Recall	F1-score
Count Vectorizer	87,14%	85,89%	87,63%	86,72%
TF-IDF	88,77%	87,45%	88,77%	87,81%

The results of this study show a comparison of the predicted values of the Support Vector Machine algorithm using the TF-IDF feature and the Count Vectorizer feature.

The results show that the percentage of hate speech prediction using the TF-IDF feature is greater than using the Count Vectorizer feature. The accuracy obtained using the TF-IDF feature has an Accuracy Rate of 88.77%, Precision of 87.45%, Recall of 88.77%, and F1-Score of 87.81%.

IV. CONCLUSION

Based on the results of the tests that have been carried out, the authors can draw the following conclusions:

- The Support Vector Machine algorithm using the TF-IDF feature has a greater accuracy value than using the Count Vectorizer feature, which is equal to 88.77% in detecting hate speech.
- Overall, by using the TF-IDF feature, the Support Vector Machine algorithm gets high results compared to the Count Vectorizer feature, with an Accuracy Value of 88.77%, 87.45% Precision, 88.77% Recall, and F1-score of 87.81%.

REFERENCES

Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications, 41(4), 1498-1508. https://doi.org/10.1016/j.eswa.2013.08.047

Amri, A. (2020). Implementasi Algoritma Random Forest Untuk Mendeteksi Hate Speech Dan Abusive Language Pada Twitter Bahasa Indonesia (Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim Riau).

Istaiteh, O., Al-Omoush, R., & Tedmori, S. (2020, October). Racist and sexist hate speech detection: Literature review. In 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) (pp. 95-99). IEEE., doi: 10.1109/IDSTA50958.2020.9264052

Ivan, I., Sari, Y. A., & Adikara, P. P. (2019). Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 3(5), 4914-4922.

Modi, S. (2018, December). AHTDT-Automatic Hate Text Detection Techniques in Social Media. In 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET) (pp. 1-3). IEEE.

Moh, M., Moh, T. S., & Khieu, B. (2020, January). No" Love" Lost: Defending Hate Speech Detection Models Against Adversaries. In 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-6). IEEE.

Mustafa, M. S., Ramadhan, M. R., & Thenata, A. P. (2018). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Creative Information Technology Journal, 4(2), 151-162.

Rachmah, E. N., & Baharuddin, F. (2019). Faktor pembentuk perilaku body shaming di media sosial. In Prosiding Seminar Nasional & Call Paper Psikologi Sosial (pp. 66-73). Retrieved from http://fppsi.um.ac.id/wp-content/uploads/2019/07/Eva-Nur.pdf

Ritonga, A. S., & Purwaningsih, E. S. (2018). Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding). Jurnal Ilmiah Edutic: Pendidikan dan Informatika, 5(1), 17-25.

Salsabila, N. A., Winatmoko, Y. A., Septiandri, A. A., & Jamal, A. (2018, November). Colloquial indonesian lexicon. In 2018 International Conference on Asian Language Processing (IALP) (pp. 226-229). IEEE. https://doi.org/10.1109/IALP.2018.8629151

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM). Jurnal Media Informatika Budidarma, 4(3), 650-658.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access, 6, 13825-13835.
Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web, 10(5), 925-945.