Indonesian SMS Spam Detection Using TF-RF Feature Weighting Method and Support Vector Machine Classifier

Muhammad Syulhan Al Ghofany[1], Ramaditia Dwiyansaputra[1](✉), Fitri Bimantoro[1], and Khairunnas[2]

[1] Department Informatics Engineering, University of Mataram, Mataram NTB, Indonesia
rama@unram.ac.id

[2] Computer Engineering Department, Vistula University, Warzawa, Poland

Abstract. SMS Spam is an unsolicited or unwanted text message by a user that is sent to a mobile device. At this time, increasingly criminal acts can annoy recipients by spreading unsolicited or unwanted spam SMS, including promotions, fraud, pornographic messages, and others. Therefore, the classification of SMS needs to be developed to assist in categorizing SMS. In existing research, to try to overcome these problems, the term frequency-inverse document frequency (TF-IDF) feature is applied. However, this method has a disadvantage, namely eliminating category information on each document, so in this study, a comparison will be made with the Supervised Term Weighting feature method, which is one of the terms frequency relevance frequency (TF-RF) using the Support Vector Machine, K-nearest Neighbor, and Multinomial Naïve Bayes. The total data used is 500 SMS with a comparison of 325 non-spam SMS and 175 spam SMS. After the experiment is conducted, SVM Kernel Sigmoid has the highest average accuracy value where the difference in average accuracy with Kernel RBF is 2.26%, Linear Kernel is 0.09%, k-Nearest Neighbor is 27.56%, and Multinomial Naïve Bayes is 4.37%.

Keywords: SMS Spam · Text Classification · Supervised Term Weighting · TF-RF · Support Vector Machine

1 Introduction

The broader number of SMS users in the community is widely used by irresponsible parties to commit crimes by spreading unsolicited and unwanted SMS spam, such as promotions, fraud, pornographic messages, etc. SMS spam is unwanted or unsolicited text messages, and the sender is unknown [1]. Usually, the message contains an offer of something or even a form of fraudulent mode. According to the FCC, it is against the law to send commercial messages without permission to wireless devices, including cell phones and pagers, unless the sender first gets your authorization [2].

The government makes several efforts to resolve this problem, such as the revision of the Minister of Communication and Informatics Regulation No. 1/2009, which regulates

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 117–129, 2022.
https://doi.org/10.2991/978-94-6463-084-8_12

premium messaging services and will emphasize regulations related to SMS advertising. However, SMS advertising and spam are not much different because they send short messages to many users. However, the difference is that SMS advertising must follow the user's approval, even though it is still troubling. After all, the messages sent are still not controlled in quantity and time—delivery [3]. Efforts have also been made to register prepaid sim cards, followed by IMEI validation rules, but these spam SMS are still popping uncontrollably [4].

To overcome this problem, the classification technique will be applied to the SMS text of spam messages to distinguish messages that contain spam and letters that do not prevent spam. Classification is a process of finding a set of models or functions that describe and distinguish data classes to predict the class of objects whose class is not yet known (supervised learning) with categorical data type characteristics. The method that can be used in text classification is the Support Vector Machine (SVM) method. The SVM classification method is one of the most appropriate discriminatory methods. The SVM method provides high and stable accuracy values [5].

Research on the classification of SMS spam has been done quite a lot. Still, most of these studies use the standard or unsupervised term weighting method, which is currently the most popular, namely Term Frequency Inverse Document Frequency (TF-IDF) [6, 10]. Term Frequency is the occurrence of the frequency of occurrence of the same word in the document. Inverse Document Frequency is the number of related document collections containing specific words. TF-IDF gives high weight to terms that rarely appear throughout the document. This TF-IDF has the disadvantage of omitting category information in each document. TF-IDF only depends on the frequency of words in the document and the number (inversely) of training documents that contain this term.

In addition to research using traditional methods, there is also a Supervised term weighting (STW) method or supervised feature weighting where this method utilizes known information about membership of training documents into categories so that highly discriminatory terms in specific categories are beneficial in their emergence in the process. One of the modern methods available is the Term Frequency Relevance Frequency (TF-RF). Compared to the traditional method, which is only based on the distribution of terms/words throughout the document or prefers rare terms/words, the TF-RF method is one method that pays attention to terms/phrases that often appear in each document in each category. In several previous studies, the TF-RF method performs pretty well for text classification. It is even better than the traditional method that has been frequently and commonly used, namely TF-IDF [11, 14].

Based on the things described SMS spam detection in Indonesia can be applied using one of the STW methods, namely TF-RF and the Support Vector Machine classification method. Applying this method is expected to produce a good performance and as expected.

2 Literature Review

The research was conducted by adding a Genetic Algorithm in the attribute selection process that will be used in the SMS message classification process with the Naïve Bayes algorithm, which aims to filter and separate spam SMS and non-spam SMS. This study used five stages for data preprocessing: transform case, tokenization, stop word filter, stemming, and N-grams. This study uses the TF-IDF weighting technique. The success rate of SMS message classification using Nave Bayes with GA resulted in a better accuracy rate of 89.73% with an increase of 0.34%, which previously used Nave Bayes of 89.39% and the AUC value of 0.654 [6].

Research has been carried out by developing a system for the Support Vector Machine method, where SMS data obtained from the Kaggle database is processed first using tokenizing, word normalization, filtering, and stemming techniques. Furthermore, cross-validation tests the training data used in the classification process. The SVM algorithm can classify spam in SMS with an accuracy of 96.72% compared to Naive Bayes [7].

This research uses the TF IDF method for the case of sentiment analysis on Instagram comments. The Support Vector Machine (SVM) with TF IDF would be better. With the composition of the best data for testing is 80%: 20% (train data: test data), the results obtained are 87.45% accuracy, 87.72% precision, 91.74% recall, and 89.69% F1-Score in Decision Tree with TF-IDF, while for Support Vector Machine with TF IDF the best data composition for testing is 80%: 20% (train data: test data) with 94.36% accuracy, 96.78% precision, 94.30% recall and 95.53% F1 Score. Judging from the Support Vector Machine (SVM) results with TFIDF, it is better than the Decision Tree with TFIDF. Still, the results of the two algorithms are already excellent because they have accuracy above 80% [8].

Research on article classification based on the age level of the reader is carried out by applying the term frequency, inverse document frequency (TFIDF) features, and the Multinomial Naive Bayes Classifier algorithm. In this study, the article data used were sourced from 3 sites, namely, bobo.grid.id, a site with a target audience of elementary school children, for teenagers aged 15–24 years. It was obtained from the hai.grid.id site, while for the group category adult age obtained from www.detik.com. The results obtained in this study are 93% accuracy, 94% precision, and 93% recall [9].

Research has been done on the study of term weighting for text categorization by trying supervised variants of the IDF TF, namely RF TF and others. Extensive experimental studies were conducted on two datasets, the Reuters corpus with 10 or 52 categories and 20 Newsgroups, and three different classification methods, namely the SVM classifier with linear kernel function and RBF and Random Forest. The results obtained showed that the RF TF weighting method obtained better results on all datasets and with all classifiers compared to IDF TF, as in Reuters-10, sequentially with linear kernel, RBF, and Random Forest obtained an accuracy of 89%, 90%, and 85% compared to TF IDF obtained 87%, 80%, and 84% accuracy. The statistical significance test shows that the proposed scheme consistently achieves higher effectiveness and is never worse than the IDF TF method. The TF RF weighting method gives excellent results with few features and shows some decrease (less than 4%) as the number of elements increases [11].

There is a study that proposes a comparison of several term weighting methods on the results of text classification in the Hadith Translation Dataset, namely Term Frequency Inverse Document Frequency (TF-IDF), Term Frequency Inverse Document Frequency Inverse Class Frequency (TF-IDFICF), Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency (TF-IDFICSδF), and Term Frequency Inverse Document Frequency Inverse Class Space Density Frequency Inverse Hadith Space

Density Frequency (TF-IDF-ICSδF-IHSδF). This study compares the results of term weighting to the Translation of 9 Books of Hadith dataset, which is applied to the Naive Bayes and SVM classification engines. The trial results show that the classification results using the TFIDF-ICSδFIHSδF term weighting method outperform another term weighting, namely getting 90% precision, 93% recall, 92% f1-Score, and 83% accuracy [12].

A comparative study of several weighting methods for classifying news topics using a decision tree was conducted. The dataset is Indonesian news articles comprising 12 news categories such as politics, culture, health, education, and 360 documents. This test aims to determine the best weighting technique among TF-ABS, TF-CHI2, TF-RF, and TF-IDF. Based on the test, the results show that TF-ABS produces the highest accuracy of 82.22%, slightly different from TF-CHI2, which has an accuracy of 80.83%. The other two techniques have lower accuracy, namely TF-RF with an accuracy value of 65.56%, and the lowest was TF-IDF with a value of 50% [13].

Comparative analysis has been carried out on weighting the words TF-IDF and TF-RF on trending topics on Twitter and using the classification method from data mining. The method used is the K-Nearest Neighbor classification method. The amount of data used is 77793 tweet data obtained from Twitter media, which is manually labeled and divided into 12 categories: economy, entertainment, law, health, sports, automotive, education, politics, arts and culture, social, technology, general. Based on the results of testing the implementation of TF-IDF and TF-RF weighting against the K-Nearest neighbor classification, the highest accuracy results using k = 1 with a scenario (90–10) and the accuracy results obtained are 63.12% with a precision of 0.633 and a recall of 0.633. In this case, the comparison of TF-IDF and TF-RF using the K-Nearest Neighbor classification shows that TF-IDF is better in the confusion matrix [14].

Based on the various studies that have been described previously, it can be concluded that the TF-IDF weighting method has quite good results but still has shortcomings, so research is carried out with other forms with efforts to improve the previous method, one of which is TF-RF and get good results, even in some studies got better results than TF-IDF. The Support Vector Machine classification method also has good results when used for grouping text. Therefore, the research for detecting SMS spam in Indonesia will be carried out using the TF-RF method and the Support Vector Classifier, where the test will compare TF-RF with TF-IDF.

3 Research Methods

3.1 Dataset

The research material used in this final project is SMS data collected from personal cellphones and several other people's cellphones. With the owner's approval, the cellphone first downloads an application through the Playstore called SMS Backup and Restore. The SMS received from early 2021 until the newest one is still being accepted. The number of SMS data collected from these mobile phones is 500 SMS. Furthermore, the collected SMS is labeled (spam and non-spam) according to the source of the SMS sender, the time of sending the SMS, and the SMS content.

3.2 System Planning

The process diagram of the Indonesian SMS spam detection system with the TF-RF method using the support vector machine classification consists of several stages as shown in Fig. 1.

3.3 Data Preprocessing

At this stage, several things are done so that the data processing at the next stage can be processed correctly. There are 4 stages carried out in this process: Case Folding, Tokenizing, Stopword Removal, and Stemming.

1. Case Folding is a process that converts all words in a document or corpus into lowercase letters, so there is no ambiguity when comparing words starting with uppercase letters and lowercase letters in the same term or word.
2. Tokenizing is a process that functions to change a collection of sentences in the text into units of words or tokens.

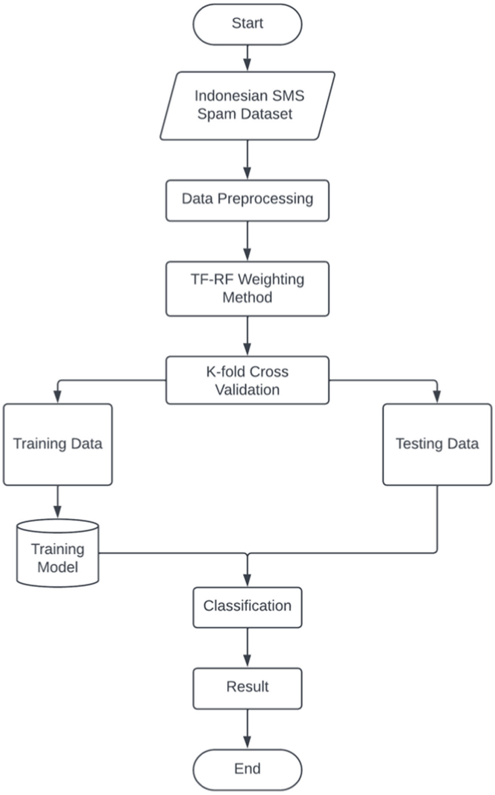

Fig. 1. Diagram of SMS Spam Classification

3. Stopword Removal is a process that serves to remove irrelevant words contained in the corpus or document. The removal of irrelevant words in the document is done so that words that do not represent the characteristics of a document are not processed so that it can speed up computing time. The stopword list used in this study was obtained from a list of conjunctions in Indonesian.
4. Stemming is a process that functions to change formed words or words with affixes into essential words by removing the prefix and suffix of the word. The stemming process is needed so that every word with different affixes is still considered the same according to the root word. The stemming algorithm used in this study is the Nazief & Adriani Algorithm obtained from the library.

3.4 Term Weighting

This stage aims to determine the pattern or the characteristics of each SMS category by weighting the terms contained in the SMS. A training dataset is also carried out in this process, which will produce a training model. The following are some of the steps carried out at the pattern discovery stage:

Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is the frequency of occurrence of a word/term t in each document d in each class. While (*t*) looks for the value of the occurrence of the term in a collection of documents, in other words, it pays attention to the number of documents d that has the word or term t. After finding these two values, the TF-IDF will be weighted by multiplying the TF value with the IDF value. The calculation of the IDF and TF-IDF values can be seen in Eqs. (1) and (2).

$$IDF(t) = log\frac{n}{df(t)} \tag{1}$$

$$TF.IDF = TF(d, t) * IDF(t) \tag{2}$$

Term Frequency-Relevance Frequency (TF-RF). Term Frequency-Relevance Frequency (TF-RF) Relevance frequency is a method that emerged to improve the existing methods. For example, the IDF method will only assess terms based on the terms' occurrence (presence or absence) in a document. In contrast to the RF method proposed by Man Lan, this method considers the document's relevance regarding the frequency of occurrence of terms in related categories [15]. Equations for calculating RF and TF-RF can be seen in Eqs. (3) and (4).

$$RF\left(t_j, c_i\right) = \log(2 + \frac{n_{ij}}{max(1, n_{\sim ij})}) \tag{3}$$

$$TF.RF = TF(d, t) * RF(t_j, c_i) \tag{4}$$

3.5 Support Vector Machine Classifier

In this research, the classification process will use the Support Vector Machine classification method. The characteristic of SMS is that it has a maximum length of 160

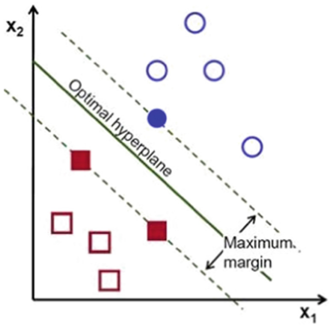

Fig. 2. Support Vector Machine Hyperplane Illustration

characters due to the limited ability of the channel used, so there are only a few that can be used as features in each message for the classification process. Due to the limited length of SMS characters available, SMS users typically use a privileged subset of language with abbreviations, phonetic contractions, poor punctuation, emoticons, etc., as opposed to more traditional languages such as those used in email. When compared to emails, spam filtering can be improved by including contextual information found in email headers, but SMS contains much less information in headers, which offers less context to work with [16]. So, for that reason, the SMS part processed from preprocessing to classification is only the subject/content part of the message. Based on the characteristics described, a classifier recommendation is the Support Vector Machine [17, 18]. Moreover, the working principle of SVM is an excellent classifier to handle the classification of 2 classes. The number of datasets to be processed is not large-scale, so this method is considered to get good results in the classification process later [19].

In simple terms, the concept of SVM is a process to find the best hyperplane which can separate two classes in the input space. The classification process can be interpreted as the process of finding the boundary line (hyperplane) that distinguishes or separates the two classes. Support Vector Machine is a classification method that aims to find the Maximum Marginal Hyperplane or MMH, which is the best dividing limit or maximum separator for all classes. Support Vector Machine (SVM) is usually used in the case of classification and regression as a prediction technique. Unlike the neural network strategy, which only tries to find the dividing hyperplane between classes, SVM will find the best hyperplane among other hyperplanes in the input space.

The best separating hyperplane between the two classes is to calculate the hyperplane margin value and look for the point with the highest value. Margin is the distance between the hyperplane and the closest pattern of each class. This closest pattern is called a Support Vector. Figure 2 shows that the line between the two classes is the best hyperplane, and the red square and blue circle are the Support Vector. The larger the margin, the higher the accuracy. Finding this hyperplane's location is the core of the learning process in SVM [20].

The way the SVM algorithm works is to describe the data set in the form of a graph (*X*i, i), where i is a collection of tuples with class labels on i. Each class can choose one of two values, namely between +1 or −1 (*y*i {+1, 1}). Getting the separating hyperplane can be done with the equation below:

$$W.X + b = 0 \tag{5}$$

If b is an additional weight, then the value can be defined as 0, as in the equation below:

$$w_0 + w_1x_1 + w_2x_2 = 0 \tag{6}$$

Thus, the value above the separating hyperplane will satisfy the following equation:

$$w_0 + w_1x_1 + w_2x_2 > 0 \tag{7}$$

The value under the separating hyperplane will satisfy the following equation:

$$w_0 + w_1x_1 + w_2x_2 < 0 \tag{8}$$

The weight of each equation can be adjusted so that the dividing hyperplane for each side can be written as the equation below:

$$H_1 : w_0 + w_1x_1 + w_2x_2 \geq +1 untuky_i = +1 \tag{9}$$

$$H_2 : w_0 + w_1x_1 + w_2x_2 \leq -1 untuky_i = -1 \tag{10}$$

From the two inequalities above, it can be seen that every tuple or value right or above 1 belongs to class +1 and every tuple or value suitable or below 2 belongs to class −1.

3.6 Evaluation

The classification results are then evaluated to obtain an accuracy value that will analyze whether the classification model is feasible [21]. The technique used to evaluate the classification in this study is to calculate recall, precision, and f-measure. This technique uses a confusion matrix as a calculation reference [22]. A confusion matrix is an accuracy calculation method that is usually used in the concept of Data Mining or Decision Support Systems. In measuring performance using the Confusion Matrix, 4 (four) represent the results of the classification process. The four terms are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The True Negative (TN) value is the number of harmful data correctly detected, while the False Positive (FP) is harmful data detected as positive data. Meanwhile, True Positive (TP) is positive data detected correctly. False Negative (FN) is the opposite of True Positive, so the data is positive but detected as unfavorable.

The confusion matrix shows the level of accuracy of the classification model that has been done previously. Accuracy shows the proportion of the number of correct predictions. The following is a formula for testing accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

Recall or true positive rate (TP) is the proposition of positive cases identified correctly, along with the formula for finding Recall.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Precision is the level of accuracy of the proportion of positive cases that have been correctly predicted (TP) with the total number of cases that have been predicted, following the formula to find Precision.

$$Precision = \frac{TP}{TPFP} \tag{13}$$

4 Result and Discussion

The testing process will be divided into 2 parts, namely testing with a dataset that uses a stemming process and a dataset that does not use a stemming process. This experiment was conducted to find out if, with the stemming process, how the accuracy results obtained from each classifier will be compared without the stemming process, namely with the Support Vector Machine with 3 kernels (Linear, RBF, and Sigmoid), k-Nearest Neighbor, and Multinomial Naïve Bayes, using either the TF-IDF or TF-RF weighting methods. Based on testing using datasets collected in this study, the results obtained are the values of accuracy, precision, and recall presented in the table for each classification method and a comparison of each method, both the weighting method and the classification method.

The test is carried out through a 10-fold cross-validation scheme which is carried out for 15 iterations. After the iteration process, the average recall, precision, and accuracy values are obtained for every 10-fold cross-validation.

Based on the results of testing using several test schemes using different weighting and classification methods as well as the use of stemming and without stemming, several average values of precision, recall, and accuracy was obtained from each experiment. Figures 3, 4 and 5 shows a comparison diagram of the average value of the 10-fold cross-validation with 15 iterations of test results on each test scheme with and without stemming.

Based on the diagram, it is known that the values of precision, recall, and accuracy has a higher tendency to be obtained through the stemming process. Based on the accuracy results in each of the above methods, it is known for the SVM method in each kernel when compared that SVM with linear and sigmoid kernels has the highest accuracy value with the same accuracy, precision, and recall values. The RBF kernel is a function used when the data is not linearly separated. At the same time, the Linear Kernel is a good kernel function to use when the data is linearly separated, so it can be concluded that the dataset has been divided linearly so that the accuracy obtained is higher with the Linear Kernel. The two kernels also have the highest accuracy among other classification methods, namely k-Nearest Neighbor and Multinomial Naïve Bayes.

Furthermore, the k-Nearest Neighbor method's best k value is k = 1. The accuracy results obtained even with the best k value are still relatively low compared to SVM,

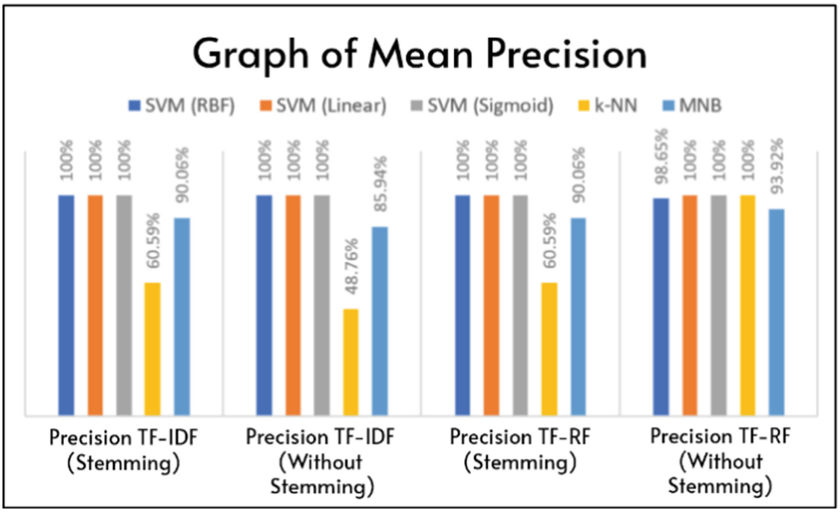

Fig. 3. Precision value comparison chart

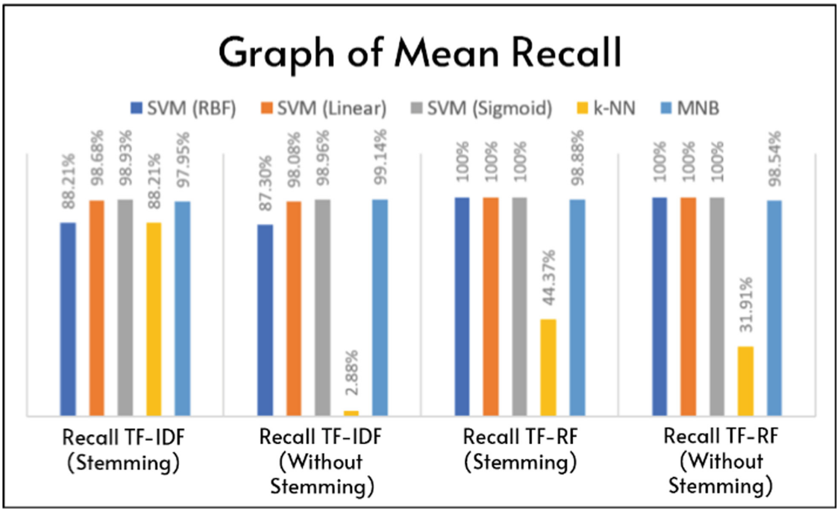

Fig. 4. Recall value comparison chart

although there is an increase when using the TF-RF weighting method. According to related studies, this could be because this method only works better on non-linear datasets and large amounts of data. In contrast to the Multinomial Naïve Bayes method, the accuracy results obtained are very high, competitive with SVM but still better than SVM for each weighting method. According to related studies, because this method does not require much training data and the process of calculating the probability value for each word is carried out, this process will produce a word in each document that characterizes

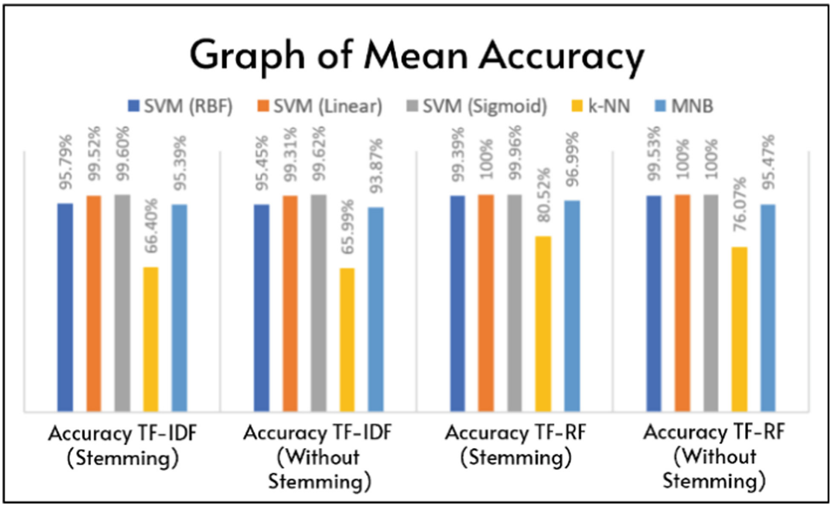

Fig. 5. Accuracy value comparison chart

documents in a specific category so that the data training process is carried out optimally, before classification like SVM.

A better weighting method between TFIDF and TF-RF is based on the results of the accuracy of each method, namely TF-RF. Although both focus on the occurrence of words, TF-RF also considers the occurrence of words based on specific categories, unlike TF-IDF which does not look at What category the document is in so that the accuracy value for TF-RF is always higher than TFIDF.

5 Conclusions

This research uses the TF-RF feature weighting method and Support Vector Machine classifier to build a model for Indonesian SMS spam detection. This research also compared the performance of the proposed method with the model commonly used for Indonesian SMS Spam detection, such as kNN and Multinomial Naïve Bayes.

Based on the research results, it can be concluded that the TFRF method has a higher average performance in terms of precision, recall, and accuracy than the TFIDF combined with each classifier method with an average difference of 5.78% precision, 1.54% recall, and accuracy of 3.7%. The Support Vector Machine classifier with sigmoid kernel method had the highest average accuracy for each weighting method with stemming and without stemming compared to each classification method, where the difference in average accuracy with RBF Kernel is 2.26%, Linear Kernel is 2.26%, 0.09%, k-Nearest Neighbor at 27.56%, and Multinomial Nave Bayes at 4.37%. The testing schemes that go through the stemming stage tend to get higher results than those that do not go through the stemming stage, and this is because the affixes for each syllable add to various features during weighting processing, thus giving different weight values.

References

1. N. Zakiah: Cara Menyingkirkan SMS Spam, supaya Gak Merasa Terganggu Lagi. https://www.idntimes.com/tech/trend/nena-zakiah-1/cara-stop-sms-spam/4, 2020
2. Okezone: Apa Itu SMS Spam?. https://techno.okezone.com/read/2020/01/25/207/2158113/apa-itu-sms-spam, 2020
3. Kominfo: SMS Spam Dilarang, SMS Iklan Buka Peluang. https://kominfo.go.id/content/detail/1825/sms-spam-dilarang-sms-iklan-buka-peluang/0/sorotan_media, 2012
4. CNN: Banyak SMS Sampah, Ombudsman Kritik Registrasi Prabayar," https://www.cnnindonesia.com/teknologi/20190815135828-185-421609/banyak-sms-sampah-ombudsman-kritik-registrasi-prabayar, 2019
5. M. Imelda A. Muis & Muhammad Affandes: Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet. Sains, Teknol. dan Ind. Sultan Syarif Kasim Riau, vol. 12, no. 2, pp. 189–197 (2015)
6. I. Munitasri, S. Santosa, and C. Supriyanto: Klasifikasi Pesan Sms Menggunakan Algoritma Naive Bayes Dengan Seleksi Fitur Genetic Algorithm. J. Teknol. Inf., vol. 14, no. 1, (2018)
7. Sandag, G.A., Sambur, R.J., Bororing, J.: Klasifikasi Sms Spam Menggunakan Support Vector Machine. J. Pilar Nusa Mandiri **15**(2), 275–280 (2019). https://doi.org/10.33480/pilar.v15i2.693
8. Asshiddiqi, M.F., Lhaksmana, K.M.: Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI. Univ. Telkom **5**(3), 177–178 (2020)
9. I. D. Putra: Klasifikasi Artikel Berdasarkan Tingkatan Umur Pembaca menggunakan Metode Multinomial Naive Bayes Classifier. no. November (2019)
10. R. Dwiyansaputra, G. S. Nugraha, F. Bimantoro, and A. Aranta: Deteksi SMS Spam Berbahasa Indonesia Menggunakan TF-IDF dan Stochastic Gradient Descent Classifier (Indonesian SMS Spam Detection using TF-IDF and Stochastic Gradient Descent. J. Teknol. Informasi, Komput. dan Apl., vol. 3, no. 2, pp. 200–207 (2021)
11. G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori: A study on term weighting for text categorization: A novel supervised variant of tf.idf. DATA 2015 - 4th Int. Conf. Data Manag. Technol. Appl. Proc., pp. 26–37 (2015). https://doi.org/10.5220/0005511900260037
12. A. T. Ni'mah and A. Z. Arifin: Perbandingan Metode Term Weighting terhadap Hasil Klasifikasi Teks pada Dataset Terjemahan Kitab Hadis. Rekayasa, vol. 13, no. 2, pp. 172–180 (2020). https://doi.org/10.21107/rekayasa.v13i2.6412
13. H. Tantyoko, Adiwijaya, and U. N. Wisesty: Perbandingan Pembobotan untuk Klasifikasi Topik Berita menggunakan Decision Tree. J. Teknol. APERTI BUMN, vol. 2, pp. 97–113, (2019)
14. Assidyk, A.N., et al.: Analisis Perbandingan Pembobotan TF-IDF dan TF-RF pada Trending Topic di Twitter dengan Menggunakan Klasifikasi K-Nearest Neighbor. Univ. Telkom **7**(2), 7773–7781 (2020)
15. Lan, M.: A New Term Weighting Method for Text Categorization. Natl. Univ, Singapore (2006)
16. Delany, S.J., Buckley, M., Greene, D.: SMS spam filtering: Methods and data. Expert Syst. Appl. **39**(10), 9899–9908 (2012). https://doi.org/10.1016/j.eswa.2012.02.053
17. Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., Odusami, M.: A review of soft techniques for SMS spam classification: Methods, approaches and applications. Eng. Appl. Artif. Intell. **86**(August), 197–212 (2019). https://doi.org/10.1016/j.engappai.2019.08.024
18. P. Navaney, G. Dubey, and A. Rana: SMS Spam Filtering Using Supervised Machine Learning Algorithms. Proc. 8th Int. Conf. Conflu. 2018 Cloud Comput. Data Sci. Eng. Conflu. 2018, pp. 43–48 (2018). https://doi.org/10.1109/CONFLUENCE.2018.8442564

19. Kadhim, A.I.: Survey on supervised machine learning techniques for automatic text classification. Artif. Intell. Rev. **52**(1), 273–292 (2019). https://doi.org/10.1007/s10462-018-09677-1
20. N. S. Dhuha: Klasifikasi Teks Pengaduan Sambat Online Menggunakan Support Vector Machine (SVM). J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 4 (2020)
21. H. N. Irmanda and R. Astriratma: Klasifikasi Jenis Pantun dengan Metode Support Vector Machines (SVM). RESTI, vol. 1, no. 10 (2021)
22. A. Sabrani, I. G. Putu, W. Wedashwara, and F. Bimantoro: Metode Multinomial Naïve Bayes untuk Klasifikasi Artikel Online Tentang Gempa di Indonesia (Multinomial Naïve Bayes Method for Classification of Online Article About Earthquake in Indonesia). vol. 2, no. 1, pp. 89–100 (2020)