ETLE Sentiment Analysis Performance Increasement with TF-IDF, MDI Feature Selection, and SVM

[1] Muhammad Syiarul Amrullah*, [2]Aji Gautama Putrada, [3]Mohamad Nurkamal Fauzan, [4]Nur Alamsyah
[1]Universitas Logistik dan Bisnis International
[2,3]Advanced and Creative Networks Research Center, Telkom University
[4]Universitas Informatika dan Bisnis Indonesia
[1]Jl. Sari Asih No. 54 Bandung, Jawa Barat, Indonesia
[2,3]Jl. Telekomunikasi No. 1, Bandung, Jawa Barat, Indonesia
[4]Jl. Soekarno Hatta No. 643 Bandung, Jawa Barat, Indonesia
*e-mail: syiarul45@gmail.com

*(**received:** 2 February 2023, **revised:** 28 April 2024, **accepted:** 22 June 2024)*

Abstrak

Di Indonesia, pemerintah melalui Kepolisian Republik Indonesia (POLRI) baru saja merilis aturan baru, yaitu Electronic Traffic Law Enforcement (ETLE), kebijakan tilang lalu lintas yang dilakukan secara elektronik melalui pemantauan kamera yang terhubung langsung dengan database surat tanda nomor kendaraan (STNK) masing-masing pengemudi. Pemerintah dapat mengukur kesukaan atau ketidaksukaan masyarakat terhadap kebijakan public tersebut melalui sentiment analysis. Sudah ada penelitian yang telah menerapkan sentiment analysis untuk mengetahui tanggapan masyarakat terhadap ETLE. Namun dari segi kinerja, model tersebut hanya mempunyai akurasi 0.42. Penelitian ini mengusulkan penggunaan support vector machine (SVM), term frequency-inversed document frequency (TF-IDF), dan mean decrease in impurity (MDI) untuk evaluasi polarisasi sentiment analysis pada kebijakan ETLE. Pertama kami mengambil data tweet mengenai ETLE dari twitter. Kemudian kami melakukan text analysis pre-processing dan remove stop words proses. Langkah selanjut nya adalah melakukan proses TF-IDF. Kami menerapkan dua metode feature selection untuk kami bandingkan, yaitu MDI dan recurrent feature elimination (RFE). Selanjutnya kami membandingkan dua model klasifikasi, yaitu naïve Bayes dan SVM. Beberapa metrik yang kami gunakan untuk mengevaluasi tahap pre-processing adalah probability density function (PDF) dan t-test. Sedangkan untuk mengevaluasi tahap remove stop words kami menggunakan bag of words (BoW). Finally, sensitivity, specificity, dan receiver operating curve (ROC) adalah untuk mengevaluasi metode seleksi fitur dan metode klasifikasi. Hasil pengujian memperlihatkan bahwa TF-IDF menghasilkan 1,022 fitur baru. Kombinasi dari metode-metode yang kami gunakan menghasilkan enam model yang kami bandingkan. SVM+TF-IDF+MDI adalah model dengan kinerja terbaik dibandingkan lima model lain nya. Accuracy dan area under curve (AUC) score nya berturut-turut adalah 0.99 dan 0.97.

Kata kunci: electronic traffic law enforcement, sentiment analysis, support vector machine, term frequency-inversed document frequency, mean decrease in impurity.

Abstract

In Indonesia, the government, through the Indonesian National Police (POLRI), has just released a new regulation, the Electronic Traffic Law Enforcement (ETLE). A traffic ticket policy is carried out electronically through camera monitoring connected directly to the vehicle registration certificates (STNK) database. The government can measure people's likes or dislikes of these public policies through sentiment analysis. There have been studies that have applied sentiment analysis to find out people's responses to ETLE. However, in terms of performance, this model only has an accuracy of 0.42. This study proposes the use of a support vector machine (SVM), term frequency-inversed document frequency (TF-IDF), and mean decrease in impurity (MDI) to evaluate polarization sentiment analysis on ETLE policies. First, we retrieve tweets about ETLE from Twitter. Then we do *text analysis pre-processing and the remove stop words process. The next step is to carry out the TF-IDF process. We apply two feature selection methods for our comparison: MDI and recurrent feature elimination (RFE). Next, we compare two classification models, namely naïve Bayes and SVM. Some of the metrics that we use to evaluate the pre-processing stage are the probability density function (PDF) and the t-test. We use the bag of words (BoW) to evaluate the remove stop words stage. Finally, sensitivity, specificity, and the receiver operating curve (ROC) are for evaluating feature selection methods and classification methods. The test results show that TF-IDF produces 1,022 new features. The combination of the methods we used resulted in the six models we compared. SVM+TF-IDF+MDI is the model with the best performance compared to the other five models. Accuracy and area under curve (AUC) scores are 0.99 and 0.97, respectively.*

Keywords: *electronic traffic law enforcement, sentiment analysis, support vector machine, term frequency-inversed document frequency, mean decrease in impurity.*

1 Introduction

In Indonesia, the government, through the Indonesian National Police (POLRI), has just released a new regulation, Electronic Traffic Law Enforcement (ETLE). A traffic ticket policy is carried out electronically through camera monitoring connected directly to the vehicle registration certificates (STNK) database [1]. The National Police expects several things from implementing ETLE, including the ticketing process's efficiency and effectiveness and reduced ticketing bribery [2]. The government can measure people's likes or dislikes of these public policies through sentiment analysis [3].

There have been studies that have applied sentiment analysis to find out people's responses to ETLE. Khalida *et al.* [4] used Naïve Bayes for predictive sentiment analysis on Twitter comments towards ETLE. Nevertheless, in terms of performance, this model only has an accuracy of 0.42. At the same time, Rahat *et al.* [5] proved that for sentiment analysis in the field of COVID-19, a support vector machine (SVM) can outperform the performance of naïve Bayes. There is a research opportunity to improve the performance of the sentiment analysis model on ETLE by trying SVM.

Term frequency-inversed document frequency (TF-IDF) is a feature extraction stage in sentiment analysis and is an important stage [6]. The TF-IDF process usually produces many new features in the dataset and causes the curse of dimensionality problems [7]. The feature selection process can reduce the number of these features. For example, Nafis *et al.* [8] used SVM with recursive feature elimination (RFE) to reduce the number of features in their sentiment analysis. Mean decrease of impurity (MDI) is also a feature selection method, where several studies also use MDI for text analysis. For example, Rabby *et al.* [9] used MDI to see which words are the most important in classifying documents related to COVID-19. Using MDI for feature selection on TF-IDF in the sentiment analysis process is a research opportunity.

This study proposes using SVM, TF-IDF, and MDI to evaluate polarization sentiment analysis on ETLE policies. First, we retrieve tweets about ETLE from Twitter. Then we do text analysis pre-processing and remove process stop words. The next step is to carry out the TF-IDF process. We apply two feature selection methods for our comparison: MDI and RFE. Next, we compare two classification models, naïve Bayes and SVM. Some of the metrics that we use to evaluate the pre-processing stage are the probability density function (PDF) and the t-test. Meanwhile, we use a bag of words (BoW) to evaluate the remove stop words stage. Finally, sensitivity, specificity, and the receiver operating curve (ROC) are for evaluating feature selection methods and classification methods.

To the best of our knowledge, there has never been an evaluation of sentiment analysis on ETLE policies using TF-IDF, MDI, and SVM. Here are some of our research contributions:

1. A novel sentiment analysis model for ETLE with extracted features from tweets using TF-IDF
2. Novel features for sentiment analysis model on ETLE that are selected using MDI
3. An enhanced classification model of sentiment analysis on tweet commentaries related to ETLE policies using SVM

The remainder of this paper has the arrangement as follows: Chapter 2 explains the papers related to our research and how our research addresses gaps in existing research. Chapter 3 presents the research methodology roadmap and describes each process. Chapter 4 shows the results of testing and a discussion of the contributions of this research. Finally, Chapter 5 concludes this research.

2 Literature Review

Several studies have existed regarding the application of computer science in ETLE. Pratama *et al.* [10] discussed the installation of a smart city in Jambi City, Indonesia. The installation of a smart city in the city consists of various features, including complaint applications called SIKESAL and ETLE. At the end of the study, there is a report on sentiment analysis on the feedback in SIKESAL. Khalida *et al.* [4] used Naïve Bayes for predictive sentiment analysis on Twitter comments towards ETLE. However, in terms of performance, this model only has an accuracy of 0.42.

Furthermore, several studies have used TF-IDF and SVM for sentiment analysis. Prabowo *et al.* [11] used TF-IDF and SVM for sentiment analysis on cyberbullying detection. They took their data from Instagram comments. Then the prediction results have an accuracy of 0.93. While Alkaff *et al.* [12] also used TF-IDF and SVM for sentiment analysis but for YouTube user comments on movie trailers. The best accuracy from this study is SVM, with a value of 0.86. The research opportunity is to apply TF-IDF and SVM on sentiment analysis for ETLE policies.

Several studies have implemented feature selection to increase the quality of TF-IDF. Nafis *et al.* [13] used SVM-RFE for feature selection in text analysis. This method is better than other methods and gives an accuracy of 0.98. Rabbi *et al.* [9] used MDI to see which words are the most important in classifying documents related to COVID-19. Using MDI for feature selection on TF-IDF in the sentiment analysis process is a research opportunity. Table 1 is a comparison of related works and highlights our research contribution.

Tabel 1. Related works on sentiment analysis for ETLE

Cite	ETLE	Sentiment Analysis	TF-IDF	SVM	Feature Selection	MDI
[10]	✓	✗	✗	✗	✗	✗
[4]	✓	✓	✗	✗	✗	✗
[11]	✗	✓	✓	✓	✗	✗
[12]	✗	✓	✓	✓	✗	✗
[13]	✗	✓	✓	✓	✓	✗
[9]	✗	✓	✓	✗	✓	✓
Proposed Method	✓	✓	✓	✓	✓	✓

3 Research Method

In our research, we first collect tweet data about ETLE from Twitter. Then we do text analysis pre-processing. The next step is to carry out the TF-IDF process. We apply two feature selection methods for our comparison: MDI and RFE. Next, we compare the two classification models, naïve Bayes and SVM. We use p-value, sensitivity, specificity, accuracy, and area under curve (AUC) score for evaluation. Figure 1 provides a complete description of our research methodology.

ISSN:2302-8149
e-ISSN:2540-9719

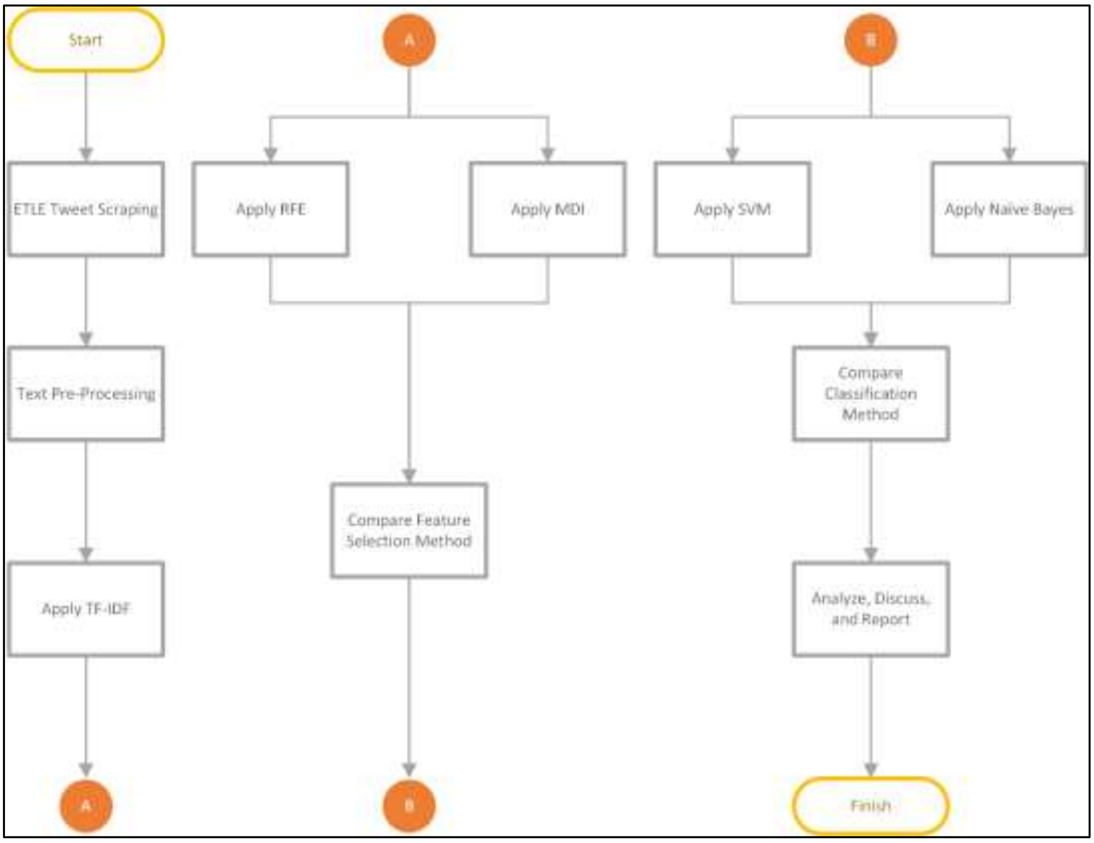


Figure 1. Our proposed research methodology

3.1 ETLE Tweet Scraping and Pre-Processing

The data retrieval stage uses the snscrape library by applying filters based on study criteria, namely tweets includes two keywords, ETLE and manual tickets, in Indonesian. The generated results consist of 691 lines of tweet data with the Comma Separated Value file format (CSVs). The file has multiple columns, including Date, Username, and Tweet.

Text pre-processing consists of several important stages [14]. The translation process translates tweets from Indonesian into English. Then, case folding makes all letters lowercase. Next, the remove stop words stage removes unnecessary words such as "and" "or," and "will." In the next step, the stemming process removes the affixes of a word. In addition, the lemmatizing process returns a word to its base word.

3.2 TF-IDF, MDI, and SVM

TF-IDF is a feature extraction method in text analysis that can indicate how important a word is in a corpus [15]. TF-IDF consists of two phases. TF describes the frequency of a word t in document d [16]. The formula for TF ($tf(t,d)$) is as equation (1) follows:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad , \qquad (1)$$

where $f_{t,d}$ describes the number of occurrences of the word t in document d and $\sum_{t' \in d} f_{t',d}$ describes the total number of words in document d.

The next process, namely IDF, describes whether a word frequently occurs in every document ($\{D|d \in D\}$) [17]. The IDF is as equation (2) follows:

ISSN:2302-8149
e-ISSN:2540-9719

$$idf(t, D) = \log \frac{N}{|\{d \in D; t \in d\}|} \quad , \qquad (2)$$

where N is the number of all documents, then $|\{d \in D; t \in d\}|$ describes the number of documents in which the word t appears.

Random forest uses MDI as its feature selection process [18]. The way to calculate MDI is to add up all the impurity reductions at each random forest node, then the results of each tree are averaged [19]. Here is the MDI formula ($Imp(X_m, T)$) is as equation (3) follows:

$$Imp(X_m, T) = \sum_{t \in T: v(s_t) = X_m} p(t) \Delta i(s_t, t) \quad , \qquad (3)$$

where T is a tree in a random forest, X_m is a feature, t is a node in T. The notation $t \in T: v(s_t) = X_m$ means every node in T whose split equation is X_m. Then $p(t)$ is the proportion of samples that arrive at node t.. Meanwhile, $\Delta i(s_t, t)$ is the decrease of impurity. The formula is as equation (4):

$$\Delta i(s_t, t) = i(t) - \frac{p(t_L)}{p(t)} i(t_L) - \frac{p(t_R)}{p(t)} i(t_R) \quad , \qquad (4)$$

where t_L is the left child of the node of a t node, t_R is a right child of a t node. Then $i(t)$ is the impurity function, and s_t is the function for splitting at t.

SVM is a supervised type of machine learning method that adds data examples to datasets with binary classes, where example data is added to the dataset to widen the distance between the two classes in the dataset [20]. If the dataset can be separated, it is linearly separable, but if it cannot, it requires adding a kernel function [21]. The kernel function transforms the dataset to a higher dimension so that the dataset appears linearly separable [22]. One of the kernel functions is the radial basis function (RBF). The RBF kernel formula ($k(x, y)$), where x is input and y is output, is as equation (5) follows:

$$k(x, y) = exp(-\gamma \|x - y\|^2) \quad , \qquad (5)$$

where $\|x - y\|^2$ is the squared Euclidean function, then γ is the independent variable.

3.3 Benchmarking and Testing Metrics

We compare the significance of the difference in tweet length before and after pre-processing with their PDF and t-test [23]. The two datasets significantly differ in tweet length if the t-test results show a $p - value < 0.05$ at a confidence level of 0.95 [24].

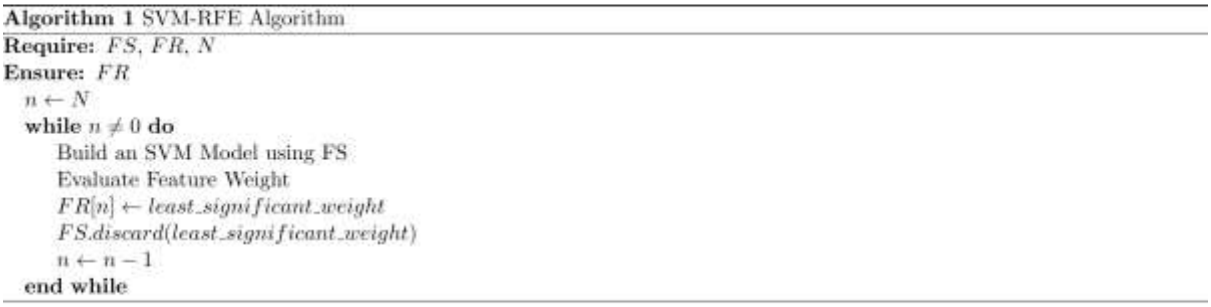
We use BoW to count the words with the highest frequency before removing stop words and after removing stop words. The BoW is a feature extraction method in which the word frequency in a sentence is calculated by assuming that a sentence is a bag of words, so no attention is given to grammar and structure [25]. The BoW formula is as equation (6) follows:

$$BoW = BoW_1 \uplus BoW_2 \uplus \cdots \uplus BoW_n \quad , \qquad (6)$$

ISSN:2302-8149
e-ISSN:2540-9719

where n is the number of documents in the corpus, then ⊎ is the disjoint union operation.

We benchmark MDI with SVM-RFE [13]. RFE is a wrapper type feature selection, where this type requires a classifier for its feature selection process, so here the classifier is SVM [26]. Figure 2 shows the SVM-RFE algorithm, where FS is a data structure that stores feature sets [27]. The feature set contains a combination of possible features. FR is a data structure that stores feature rankings. In the SVM model, feature ranking uses the weight of each feature. N is the size of the FS. The recursive process occurs in the iterative feature selection process.



```
Algorithm 1 SVM-RFE Algorithm
Require: $FS, FR, N$
Ensure: $FR$
  $n \leftarrow N$
  while $n \neq 0$ do
      Build an SVM Model using FS
      Evaluate Feature Weight
      $FR[n] \leftarrow least\_significant\_weight$
      $FS.discard(least\_significant\_weight)$
      $n \leftarrow n - 1$
  end while
```

Figure 2. The SVM-RFE algorithm

We benchmarked the SVM classification with naïve Bayes. Naïve Bayes classifies with the Bayes theorem, where there is an assumption of strong independence between features [28]. The problem that often arises in sentiment analysis cases is data imbalance. In imbalanced data, it is important to measure sensitivity and specificity [29]. In binary classification, sensitivity shows the ability of a model to predict label 1, whereas specificity, on the other hand, shows the ability of a model to predict label 0. In addition, we also calculate the accuracy of the best model. The following is the formula for sensitivity, specificity, and accuracy:

$$Sensitivity = \frac{TP}{TP + FN} \quad , \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \quad , \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad , \tag{9}$$

where TP is the true positive value, FN is the false negative value, TN is the true negative value, and FP is the false positive value [30].

ROC is a curve showing the relationship between the true positive rate (TPR) and the false positive rate (FPR) [31]. TPR is equivalent to sensitivity, while FPR has the same value as 1 – specificity [32]. AUC describes the area under the ROC curve, where a larger AUC signifies good performance, i.e. the model is able to discriminate label 0 and 1 [33]. The AUC value range is 0.5 to 1 [34]. The 0.5 value indicates that a model's predictive ability is equal to random guessing [35].

4 Results and Analysis

In this section, we present and analyze the results of our sentiment analysis on tweets related to the ETLE policies, demonstrating the effectiveness of our preprocessing steps and the comparative performance of different feature selection and classification methods.

4.1 Results

The first stage of testing is pre-processing. We analyze tweet length statistics before pre-processing and after pre-processing. The average tweet length before pre-processing is 28.3 ± 13.2 words. Meanwhile, the average tweet length after pre-processing is shorter, namely 26.2 ± 12.5

words. Through the t-test, we analyze the significance between the distributions of the two datasets. The t-test result shows that, at a confidence level = 0.95, the difference in the distribution of the two datasets is significant, with a $p - value = 0.002$. Figure 3 shows the PDF between the two datasets.

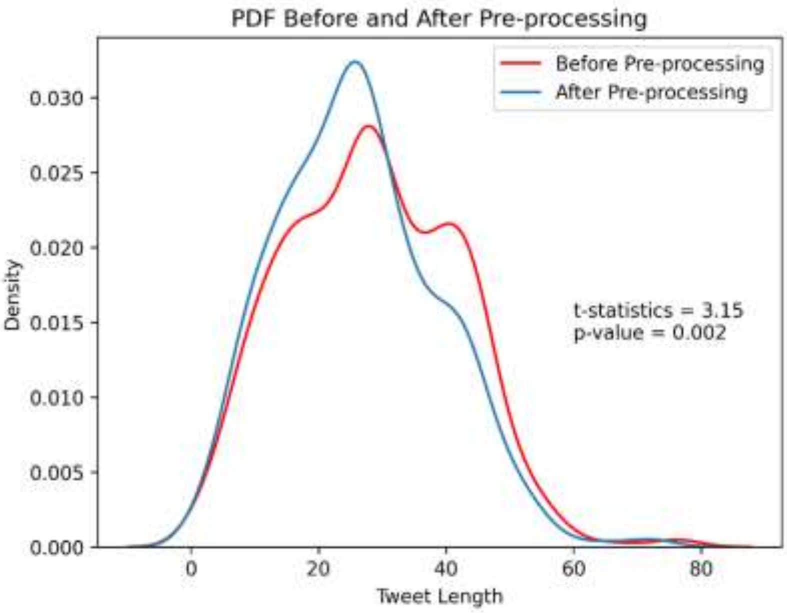


Figure 3. PDF comparison of tweet length before and after pre-processing stage

In the next step, we implement the remove stop words process. We use BoW to analyze changes in the dataset before and after removing stop words. Figure 4 shows the bar chart of BoW results on the dataset before removing stop words and after removing stop words. Before removing stop words, the two highest number words are "the," with a total of 1,232, and "of," with 622. After the remove stop words stage, these two words no longer exist. Instead, the top 5 words are "etle" with 543, traffic with 404, "police" with 401, "electronic" with 169, and "national" with 159.

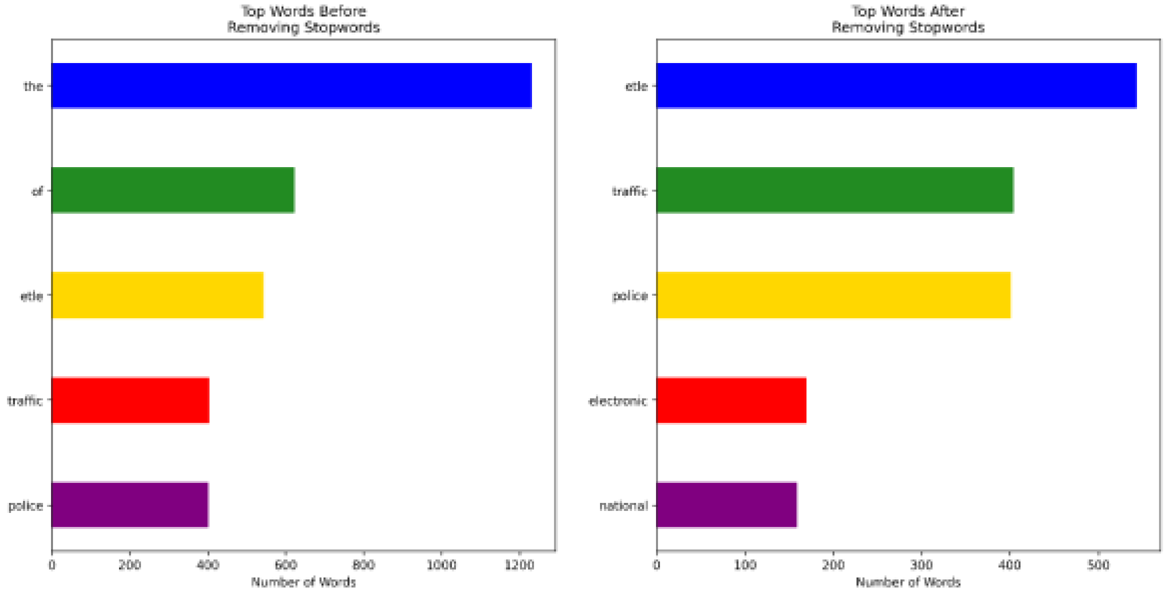


Figure 4. BoW comparison before and after removing stop words

We then implemented TF-IDF. By using two types of feature selection and two types of classification, we have the following models that we can compare:

1. Sentiment analysis with SVM and:
    a. without feature selection
    b. with SVM-RFE feature selection
    c. with MDI feature selection
2. Sentiment analysis with naïve Bayes and:
    a. without feature selection
    b. With SVM-RFE feature selection
    c. With MDI feature selection

So, we have six models. We tested each model based on its sensitivity and specificity values. TF-IDF produces 1022 new features, while SVM-RFE selects these features to 402. MDI also selected these features to be 402. Figure 5 shows a performance comparison of the six models. Five models, except the Naïve Bayes+TF-IDF model, have sensitivity = 1.0. Among the five models, SVM+TF-IDF+MDI has the highest specificity, which is 0.94. The naïve Bayes+TF-IDF model has a higher specificity, 1.0, but has a sensitivity value = 0.00. We also calculate the accuracy of the SVM+TF-IDF+MDI model. Its value is 0.99.

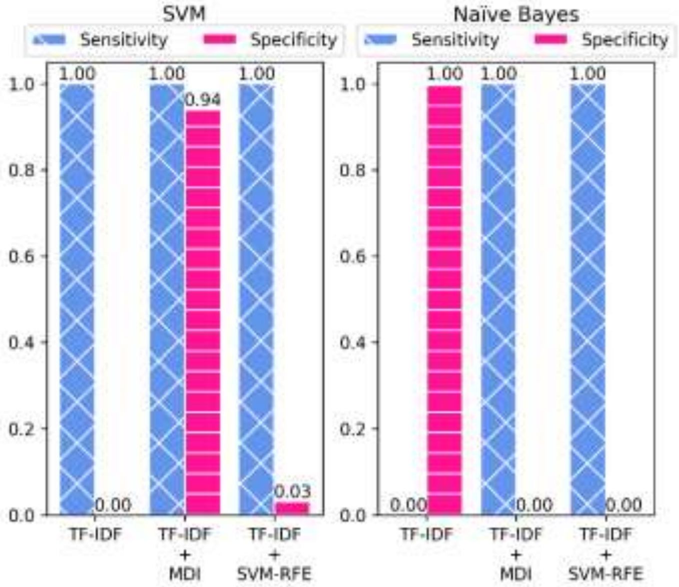
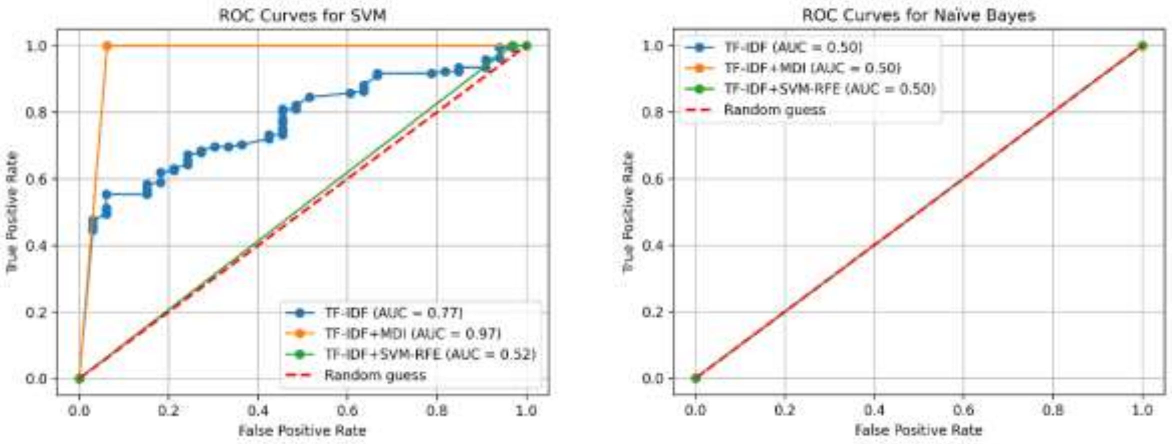


Figure 5. Sensitivity and specificity comparisons of model performance from the combination of TF-IDF, two classification methods, and two feature selection methods

To get a value that summarizes sensitivity and specificity, we test the ROC of each model. Figure 6 shows the ROC of the six models. The ROC with the highest AUC is SVM+TF-IDF+MDI, with a value of 0.97. The AUC of the other two SVM models is lower than SVM+TF-IDF+MDI but higher than the three models using naïve Bayes. All methods using naïve Bayes have AUC = 0.50, the same as a random guess.



(a) (b)

Figure 6. ROC curve of three methods, TF-IDF, TF-IDF+MDI, and TF-IDF+SVM-RFE with: (a) SVM classification (b) Naïve Bayes classification

4.2 Analysis

Paper [5] succeeded in proving that SVM works better than Naïve Bayes in the case of sentiment analysis on COVID-19 tweets. Here our contribution is the application of proof that SVM is better than Naïve Bayes in the case of sentiment analysis in tweets related to ETLE policies.

Previous research, namely paper [4], has conducted sentiment analysis on tweets related to ETLE. Nevertheless, the accuracy is only 0.42. Our contribution is a sentiment analysis model related to ETLE, which adds to the TF-IDF process. This novel model has a performance of 0.99, better than the state-of-the-art model.

Our feature extraction using TF-IDF resulted in 1,022 new features. That many features trigger the curse of dimensionality. Other studies apply SVM-RFE [8], then some apply MDI [9] to reduce dimensions with feature selection. Our contribution is to prove that MDI is better than SVM-RFE in reducing dimensions while improving sentiment analysis performance on ETLE-related tweets.

5 Conclusion

This study implements a sentiment analysis on tweets about ETLE. We use a dataset crawled from Twitter where 337 tweets have positive sentiments, while 65 tweets have negative sentiments. We apply TF-DF as a feature extraction process, compare MDI and SVM-RFE as its feature selection method, and compare SVM and naïve Bayes as its sentiment classification method. The test results show that TF-IDF produces 1,022 new features. The combination of the methods we use results in six models. The SVM+TF-IDF+MDI is the model with the best performance compared to the other five models. The Accuracy and AUC, respectively, are 0.99 and 0.97.

References

[1] E. Syafitri and D. Mashur, “Efektivitas Implementasi Program Electronic Traffic Law envorcement (ETLE) Nasional dalam Peningkatan Pelayanan Publik di Kota Pekanbaru,” Cross-Bord., vol. 5, no. 2, pp. 1322–1337, 2022.

[2] F. A. Abdullah and F. Windiyastuti, “Electronic Traffic Law Enforcement (ETLE) sebagai Digitalisasi Proses Tilang,” J. Kewarganegaraan, vol. 6, no. 2, pp. 3004–3008, 2022.

[3] E. Georgiadou, S. Angelopoulos, and H. Drake, “Big Data Analytics and international negotiations: sentiment analysis of Brexit Negotiating Outcomes,” Int. J. Inf. Manag., vol. 51, p. 102048, 2020.

[4] R. Khalida and S. Setiawati, “Analisis Sentimen Sistem E-Tilang menggunakan Algoritma Naive Bayes dengan Optimalisasi Information Gain,” J. Inform. Inf. Secur., vol. 1, no. 1, 2020.

[5] A. M. Rahat, A. Kahir, and A. K. M. Masum, “Comparison of Naive Bayes and Svm Algorithm Based on Sentiment Analysis using Review Dataset,” in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019, pp. 266–270.

[6] A. G. Putrada, I. D. Wijaya, and D. Oktaria, “Overcoming Data Imbalance Problems in Sexual Harassment Classification with SMOTE,” Int. J. Inf. Commun. Technol. IJoICT, vol. 8, no. 1, pp. 20–29, 2022.

[7] A. Madasu and S. Elango, “Efficient Feature Selection Techniques for Sentiment Analysis,” Multimed. Tools Appl., vol. 79, no. 9, pp. 6313–6335, 2020.

[8] N. S. M. Nafis and S. Awang, “An Enhanced Hybrid Feature Selection Technique using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination For Sentiment Classification,” IEEE Access, vol. 9, pp. 52177–52192, 2021.

[9] G. Rabby and P. Berka, “Multi-Class Classification of COVID-19 Documents using machine Learning Algorithms,” J. Intell. Inf. Syst., pp. 1–21, 2022.

[10] I. Pratama and S. Suswanta, “Artificial Intelligence in Realizing Smart City through City Operation Center,” in International Conference on Public Organization (ICONPO 2021), 2022, pp. 53–60.

[11] W. A. Prabowo and F. Azizah, “Sentiment Analysis for Detecting Cyberbullying using TF-IDF And SVM,” J. RESTI Rekayasa Sist. Dan Teknol. Inf., vol. 4, no. 6, pp. 1142–1148, 2020.

[12] M. Alkaff, A. R. Baskara, and Y. H. Wicaksono, “Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM,” in 2020 Fifth International Conference on Informatics and Computing (ICIC), 2020, pp. 1–5.

[13] N. S. M. Nafis and S. Awang, “The Evaluation of Accuracy Performance in An Enhanced Embedded Feature Selection for Unstructured Text Classification,” Iraqi J. Sci., pp. 3397–3407, 2020.

[14] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, “Machine Learning Methods in Smart Lighting Toward Achieving User Comfort: A Survey,” IEEE Access, vol. 10, pp. 45137–45178, 2022, doi: 10.1109/ACCESS.2022.3169765.

[15] A. Thakkar and K. Chaudhari, “Predicting Stock Trend using an Integrated Term Frequency–Inverse Document Frequency-Based Feature Weight Matrix with Neural Networks,” Appl. Soft Comput., vol. 96, p. 106684, 2020.

[16] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, “Sentiment Analysis and Classification of Indian Farmers’ Protest using Twitter Data,” Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, p. 100019, 2021.

[17] S.-W. Kim and J.-M. Gil, “Research Paper Classification Systems Based on TF-IDF and LDA schemes,” Hum.-Centric Comput. Inf. Sci., vol. 9, no. 1, pp. 1–21, 2019.

[18] E. S. Saputra, A. G. Putrada, and M. Abdurohman, “Selection of Vape Sensing Features in IoT-Based Gas Monitoring with Feature Importance Techniques,” in 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1–5.

[19] A. Sutera, G. Louppe, V. A. Huynh-Thu, L. Wehenkel, and P. Geurts, “From global to local MDI variable importances for random forests and when they are Shapley values,” Adv. Neural Inf. Process. Syst., vol. 34, pp. 3533–3543, 2021.

[20] M. Ameliasari, A. G. Putrada, and R. R. Pahlevi, “An Evaluation of Svm In Hand Gesture Detection using Imu-Based Smartwatches for Smart Lighting Control,” J. Infotel, vol. 13, no. 2, pp. 47–53, 2021.

[21] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, “CIMA: A Novel Classification-Integrated Moving Average Model for Smart Lighting Intelligent Control Based on Human Presence,” Complexity, vol. 2022, pp. 1–19, Sep. 2022, doi: 10.1155/2022/4989344.

[22] B. A. Fadillah, A. G. Putrada, and M. Abdurohman, “A Wearable Device for Enhancing Basketball Shooting Correctness with MPU6050 Sensors and Support Vector Machine Classification,” Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control, 2022.

[23] M. B. Satrio, A. G. Putrada, and M. Abdurohman, “Evaluation of Face Detection and Recognition Methods in Smart Mirror Implementation,” in Proceedings of Sixth International Congress on Information and Communication Technology, 2022, pp. 449–457.

[24] S. F. Pane, Heriyanto, A. G. Putrada, N. Alamsyah, and M. N. Fauzan, “The Influence of The COVID-19 Pandemics in Indonesia On Predicting Economic Sectors,” in 2022 Seventh International Conference on Informatics and Computing (ICIC), Dec. 2022, pp. 1–6. doi: 10.1109/ICIC56845.2022.10006897.

[25] H. J. Alyamani, “Determining Feature-Size for Text to Numeric Conversion based on BOW and TF-IDF,” IJCSNS, vol. 22, no. 1, p. 283, 2022.

[26] D. Elavarasan, D. R. Vincent PM, K. Srinivasan, and C.-Y. Chang, “A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling,” Agriculture, vol. 10, no. 9, p. 400, 2020.

[27] H. Jeon and S. Oh, “Hybrid-Recursive Feature Elimination for Efficient Feature Selection,” Appl. Sci., vol. 10, p. 3211, May 2020, doi: 10.3390/app10093211.

[28] S. Chen, G. I. Webb, L. Liu, and X. Ma, “A novel selective naïve Bayes algorithm,” Knowl.-Based Syst., vol. 192, p. 105361, 2020.

[29] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, “XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data,” in 2022 International Symposium on Electronics and Smart Devices (ISESD), Nov. 2022, pp. 1–7. doi: 10.1109/ISESD56103.2022.9980630.

[30] A. G. Putrada and D. Perdana, “Improving Thermal Camera Performance in Fever Detection during COVID-19 Protocol with Random Forest Classification,” in 2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), 2021, pp. 1–6.

[31] J. Singla, “Comparing ROC Curve Based Thresholding Methods In Online Transactions Fraud Detection System using Deep Learning,” in 2021 international conference on computing, communication, and intelligent systems (ICCCIS), 2021, pp. 9–12.

[32] A. S. Jadhav, “A novel Weighted TPR-TNR Measure To Assess Performance of The Classifiers,” Expert Syst. Appl., vol. 152, p. 113391, 2020.

[33] J. Yin, F. Mutiso, and L. Tian, “Joint Hypothesis Testing of the Area Under the Receiver Operating Characteristic Curve and The Youden Index,” Pharm. Stat., vol. 20, no. 3, pp. 657–674, 2021.

[34] A. Wubalem and M. Meten, “Landslide Susceptibility Mapping using Information Value And Logistic Regression Models In Goncha Siso Eneses Area, Northwestern Ethiopia,” SN Appl. Sci., vol. 2, pp. 1–19, 2020.

[35] J. Pereira and F. Saraiva, “Convolutional Neural Network Applied To Detect Electricity Theft: A Comparative Study on Unbalanced Data Handling Techniques,” Int. J. Electr. Power Energy Syst., vol. 131, p. 107085, 2021.