Forward feature selection for toxic speech classification using support vector machine and random forest

Agustinus Bimo Gumelar[1,3], Astri Yogatama[2], Derry Pramono Adi[1], Frismanda[3], Indar Sugiarto[4]
[1]Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Department of Communication Studies, Petra Christian University, Surabaya, Indonesia
[3]Fakultas Ilmu Komputer, Universitas Narotama, Surabaya, Indonesia
[4]Department of Electrical Engineering, Petra Christian University, Surabaya, Indonesia

Article Info

Article history:

Received Aug 31, 2021
Revised Feb 10, 2022
Accepted Feb 28, 2022

Keywords:

Feature selection
Forward selection
Random forest
Support vector machine
Toxic speech

ABSTRACT

This study describes the methods for eliminating irrelevant features in speech data to enhance toxic speech classification accuracy and reduce the complexity of the learning process. Therefore, the wrapper method is introduced to estimate the forward selection technique based on support vector machine (SVM) and random forest (RF) classifier algorithms. Eight main speech features were then extracted with derivatives consisting of 9 statistical sub-features from 72 features in the extraction process. Furthermore, Python is used to implement the classifier algorithm of 2,000 toxic data collected through the world's largest video sharing media, known as YouTube. Conclusively, this experiment shows that after the feature selection process, the classification performance using SVM and RF algorithms increases to an excellent extent. We were able to select 10 speech features out of 72 original feature sets using the forward feature selection method, with 99.5% classification accuracy using RF and 99.2% using SVM.

Corresponding Author:

Agustinus Bimo Gumelar
Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember
Surabaya, 60111, Indonesia
Email: bimogumelar@ieee.org

1. INTRODUCTION

The phenomenon of toxic speech has become an interesting topic to be discussed recently because of the way it has affected people. For example, a study conducted by Tirell described how toxic speech disassociated people into specific groups and also caused the death of over 800,000 in 1994 [1]. Nowadays, freedom of expression has contributed favorably to employing toxic speech to the extent that any outpouring of thought, especially social media, often leads to toxic speech [2]. In fact, the interaction on social media, both verbal and non-verbal, has become a current lifestyle performed by many groups, and the exchange is often interspersed by toxic speech [3], [4]. The social media platforms not only allow for improved communication, but they also allow internet users to express their thoughts, which are quickly shared with the rest of the world. Furthermore, given the users' many backgrounds, beliefs, ethnicity, and cultures on these platforms, many of them prefer to use derogatory, aggressive, and hostile language when conversing with those who do not share their background [5], [6]. The amount of hate speech and toxic content on the internet has steadily increased. Since the terms "profane," "hate," and "offensive" are used interchangeably, these have been grouped as "toxic" [7], [8].

Meanwhile, this toxic speech needs to be classified using acoustic features based on their respective roles. Therefore, choosing the suitable subset of features will be helpful to reduce the computation complexity because not all attributes are relevant to the addressed classification problem. In order to achieve this, attribute selection and dimension reduction techniques are often used [9], [10]. The general stages of the feature selection process are shown in Figure 1 [11].

Past studies related to feature selection include feature optimization, dimension efficiency, and elimination of redundant features [12]. Furthermore, they aimed to improve the accuracy of the classification process in support vector machine (SVM) [13], [14]. The implementations of this classification process include feature selection by using Gaussian mixture model [13], automatic speech assessment [15], heart sound features in the frequency and time domain [16], music genre classification [17], [18], and audio-visual recognition [19], [20].

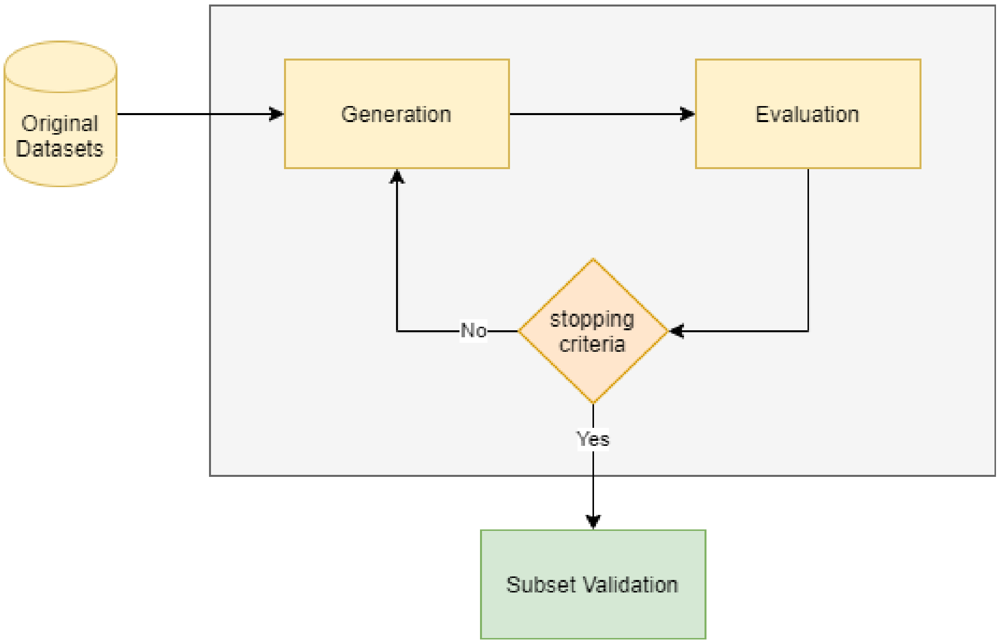

Figure 1. Stages of the feature selection process

There are three main reasons for implementing the reduction of dimensions, including minimizing the learning costs, improving the model performance, and eliminating similar or excessive dimensions [21]. This reduction is conducted because not all features are useful in the classification process. Also, some irrelevant features, known as noise, often reduce the accuracy. The deep learning algorithms perform this function by searching through the subset of possible feature spaces and evaluating each subset with quality performance. Furthermore, a sequential forward selection strategy can also be used to enhance this process. According to Ververidis and Kontropoulus [22], a simple sequential search strategy helps produce results quickly, as seen in (1), where F_1 is the cardinality of the selected features, and the other F is the original feature set.

$$O\left(F + (F-1) + (F-2) + \cdots + (F - F_1 + 1)\right) \tag{1}$$

Based on the brief introduction, toxic speech classification and feature selection are the main problems in this research. This study proposed a solution for both feature selection for toxic speech classification. We used a wrapper method with a forward feature selection technique for 2,000 data samples, and 72 features were used. Subsequently, this proposed method was applied in SVM and random forest (RF) algorithms in order to classify toxic speech. We also compared the RF result with the SVM result using selected features and the original feature set.

Moreover, this study is organized as: section 1 introduces toxic speech and how the number of speech features is decreased intentionally while still retaining the best classification result. Section 2 elucidates the method used for data preparation, speech features, and their selection method and classifier model. Furthermore, section 3 presents all results from the experiments' scenario, while Section 4 draws conclusions and interpretations from our result.

2. METHOD

A sequential search strategy was proposed using a wrapper method with a forward selection algorithm. The selected features are then used to classify toxic speech using SVM and RF. Hence, the stages in this experiment as shown in Figure 2 include data retrieval, pre-processing of both training and test data, feature extraction, feature selection, and finally, the classification and evaluation process.

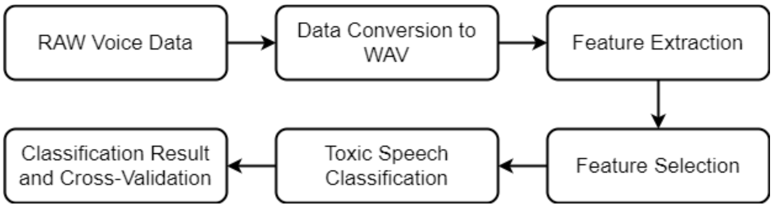

Figure 2. Experimental flow design

2.1. Dataset

The data used was collected from YouTube by using the search keywords "Online Debt" and "Online Fraud," with several video features containing conversations between two people. Consequently, 79 recordings were obtained with varying duration of 7-54 minutes. However, some of these videos contain toxic speeches with less noisy audio quality; hence the audio part was extracted and converted into a WAV file format.

2.2. Data preparation

In this section, the selection process was carried out manually by listening to the recording. The speech data was classified into toxic and non-toxic speeches manually as well. We reviewed about 2,000 speech data after selection process, with the composition of 1,273 and 763 toxic and non-toxic speeches, respectively. According to Oxford dictionary media [19] and Webster [20], toxic speech is a sentence containing intimidation, hate speech, and curses that affect the other person psychologically and emotionally. Moreover, the database used was obtained from Schofield *et al.* [23]. Eight main features were extracted from the 2,000 sound clips using openSMILE with INTERSPEECH 2010 Challenge parameterization [24]. The feature set includes energy, intensity, loudness, jitter local, JitterDDP, shimmer, harmonic noise-to-ratio, zero-crossing rate, and its nine constant statistical sub-features of maximum, minimum, range, maximum position, minimum position, mean, standard deviation, skewness, and kurtosis, respectively.

Several studies have shown, speech features obtained significant results [25]. For example, energy was used to detect a voiced or unvoiced speech, and the result showed that voiced speech has a greater energy value than unvoiced speech [26]. Similarly, sound wave intensity affects the classification of toxic sentences. By definition, the intensity of the sound wave (I) on a surface is the average rate per unit area in which energy is transferred through or to the surface. This intensity is formulated as seen in (2), where T is the energy transfer time from sound waves, and A is the surface area that cuts off the sound.

$$I = \frac{T}{A} \tag{2}$$

$$\beta = 10 \log \frac{I}{I_o} \tag{3}$$

Additionally, intensity can also be defined as sound level (β), as shown in (3), where dB stands for decibel, a unit for the level of a sound. I_o is the reference standard intensity 1,0 x 10^{-12} W/m^2 [27]. Another feature employed to classify toxic sentences is loudness, which is a volume measurement. This feature is used because whenever people are angry, they tend to make a loud voice and sometimes accompanied by curse words. However, it does not mean that every loud voice contains toxic speech.

Also, Jitter, also known as Relative Jitter, was used, and it represents the ratio of the average and sequential fundamental period. Parameter values related to the changes in the fundamental period increases due to the presence of irregular glottal vibrations [28]. Another feature used is JitterDDP, the difference in absolute average between successive periods divided by the average period [28], shown in (4).

$$JitterDDP = \frac{1}{N-1}|t_i - t_{i+1}| \tag{4}$$

$$Shimmer = \frac{1}{N-1}\sum \left|20_{log}\left(\frac{A_{i+1}}{Ai}\right)\right| \tag{5}$$

The sixth feature is the shimmer local shown in (5), and it is used to identify micro signals from the amplitude. This feature states the relative average perturbation per period of amplitude. However, this value appears and escalates in the case of organic and functional vocal pathology [28].

The seventh feature is the harmonic noise-to-ratio (HNR), which assesses the ratio between periodic and non-periodic components consisting of sound segments that are voiced [29]. The periodic component arises from the vibration of the vocal cords, whereas the non-periodic component arises from the glottal noise; meanwhile, both are usually expressed in dB. Basically, the greater the flow of air from the lungs into the vocal cords, the greater the HNR. In this case, low HNR indicates asthenic sounds and dysphonia. HNR is considered pathological when the value is less than 7 dB, as described in Boersema's study [30], and it is formulated in (6).

Even though the formula in (6) works indirectly as the definition of the frequency domain, it still produces more accurate results; hence the precision is estimated as autocorrelation $r_x(\tau) = \int x(t)x(t+\tau)dt$, where $x(t)$ is the time signal, and τ is defined as the time lag. For perfect periodic sounds, HNR has no limits [30].

$$HNR(in\, dB) = 10.^{10}\, log\frac{r_x'(\tau_{\max})}{1-r_x'(\tau_{\max})} \tag{6}$$

The zero-crossing rate (ZCR) is the last feature, which indicates the number of times the amplitude passes the 0 value in the signal data bit. Unvoiced speech often has a higher zero-crossing value than the voiced [26]. Hence, zero-crossing is detected when the data sample of a digital signal has a different sign. Consequently, the following ZCR is formulated as seen in (7), where $func$ is a function that indicates zero value when it is negative and 1 when it is positive. Furthermore, Z_w is the zero-crossing characteristic value, while N is the total number of bits contained in the frame w, and $x[m]$ is the amplitude value in the m-index data [31].

$$Z_w = \frac{1}{2}\sum_{m=1}^{N}|func(x[m] - func(x[m-1])| \tag{7}$$

2.3. Feature selection

There are two possible approaches for feature selection, namely the wrapper and filter approaches. Both are done by selecting a subset of features before the data mining algorithm is carried out. Meanwhile, the difference between the two approaches lies in the evaluation stage. For example, the wrapper approach uses the target of the selected algorithm. It then seeks in sequence, while the filter uses a separate evaluation technique from data mining algorithms for its evaluator [32].

Therefore, a wrapper method with a forward selection technique was proposed, which uses a simple search algorithm based on a linear regression model to reduce the dimensions of the dataset by eliminating redundant and irrelevant attributes [33]. Moreover, Python was used for implementing the forward selection framework originally described by Zhang [34]. Table 1 shows the original features extracted from speech data and its sub-features, while Table 2 shows the selected ones after the forward feature selection technique has been implemented. Hence, the selected features are used for the latter classification process.

Table 1. List of original speech features

No	Features	Number of Sub-features
1	Energy,	9
2	Intensity,	9
3	Loudness,	9
4	Jitter Local,	9
5	JitterDDP	9
6	Shimmer,	9
7	Harmonic Noise-to-Ratio	9
8	Zero-Crossing Rate	9
	Total Features	72

Table 2. List of selected features

No	Features	Sub-features
1	HNR	Skewness
2	Energy	Skewness
3	Loudness	Minimum
4	Loudness	Maximum Position
5	Loudness	Standard Deviation
6	Jitter Local	Minimum
7	Shimmer Local	Mean
8	Shimmer Local	Standard Deviation
9	ZCR	Minimum
10	ZCR	Skewness
Total Features		10

2.4. Classification methods

In this study, SVM and RF algorithms were used for classification. According to Pierre-Yves, SVM was a binary classifier model with a two-stage classification [35]. In the first stage, kernel functions are used to change the dimensions of features from low to high. Afterward, the non-linear data found in the highest dimension is transformed into a linear one. In the second stage, the maximum hyperplane distance was constructed to determine the decision boundary for each class [36]. Based on (8), we implement SVM using the Orange3 toolbox [37], where w_i is the Lagrange Multiplier, b is the value limit, and Z is the kernel function. The default linear kernel was used for the SVM.

$$f(x) = sign \sum_{i=1}^{L} w_i l_i Z(x_i \times x) + b \tag{8}$$

However, RF is used for the classification process in the second algorithm. According to Probst *et al.* RF algorithm depends on random vector values taken independently with the distribution in the same forest [38]. Furthermore, it utilizes ensemble learning and a prediction method with several stages of the learner. For example, Bootstrap aggregation, also known as bagging, is one of the Ensemble Learning algorithms used in RF. This method is formulated in (9), where $X = x_1, .., x_n,$ denotes a training set, $Y = y_1, .., y_n,$ denotes response, and B denotes the bagging repetition. Samples with replacement contents are $X_b, Y_b,$ whereas the amount of data is denoted by n [39]. The regression tree is denoted by f_b on $X_b, Y_b,$ and after the training process, *x'* denotes the prediction results.

$$\hat{f} = \frac{1}{W} \sum_{w=1}^{W} f_w(x') \tag{9}$$

3. RESULTS AND DISCUSSION

The selection process begins with preparing data and feature subsets, followed by toxic speech classification using SVM and RF algorithms with random samples from 2,000 data having training data and test data in the ratio of 80:20, 70:30, and 60:40. Hyperparameter of the kernel is set to Radial Basis Function, gamma is set to 1/72 (number of features), *C* parameter is set to 1, and *nu* parameter is set to 0.5, for SVM. Hyperparameter of tree number is set to 100, maximal depth of trees is set to 5, and seed is set to 1 for RF [40]. In this process, validation was done using the confusion matrix (CM) method [41]. However, before this evaluation stage with CM, cross-validation was implemented with 3-folds and 10-folds by repeatedly running the random samples of 10 times with training and test data in the ratios 60:40, 70:30, and 80:20. Initially, the data were processed by using two learning algorithms from SVM and RF. Cross-validation and random samples were obtained in order to determine the class predictions from 2,000 variables. Therefore, the results are inputted to a CM to observe the wrong variable classified by the learning machine.

3.1. Test results with random sample

The experimental result showed that the SVM and RF algorithm improves the accuracy and decreases the computing time on the entire training and test data composition when the random sample method was used. According to Table 3, the SVM classifier has an upward trend of accuracy in the ratio of all training and test data, ranging from 96.5% to 97.3%; then 96.7% to 97.3%, and 96.8% to 97.5%, for each data ratio of 80:20, 70:30, and 60:40 respectively. Meanwhile, the RF produces a less significant increase in the ratios of 80:20 and 70:30, which experienced the exact change from 94% to 95.1%, with an increase of 1.1%. Similarly, the data ratio of 60:40 received an increase of 0.2% from 93.5% to 94.7%.

Table 3. Accuracy results of random sample classification

Data Ratio (Train:Test)	All Features		Forward Selection	
	SVM	RF	SVM	RF
80:20	96.5%	94%	97.3%	95.1%
70:30	96.7%	94%	97.3%	95.1%
60:40	96.8%	93.5%	97.5%	94.7%

According to Table 4, a significant difference is seen in the computational training time of the SVM algorithm, where the data ratio of 80:20 with an initial time of 10.578 seconds becomes 3.022. Also, the 70:30 with an initial time of 6.586 seconds was processed faster to 2.56 seconds, while the 60:40 data with an entire feature duration of 6.558 seconds becomes 3.042 seconds. The RF algorithm also decreased in the computational training process. For example, the data ratio of 80:20, which takes 5.257 seconds for all features, eventually requires 0.688 after feature selection. Likewise, the 70:30, which requires 3.645 seconds, changes to 0.753 seconds, while the 60:40 with initial time, 3.885 seconds, becomes 1.169 seconds. Generally, the RF algorithm experiences a significant downward trend with an average time among all features. Hence it shows an excellent result.

Table 4. Computational time in training process of random sample (in seconds)

Data Ratio (Train:Test)	All Features (s)		Forward Selection	
	SVM	RF	SVM	RF
80:20	10.578	5.257	3.022	0.688
70:30	6.586	3.645	2.56	0.753
60:40	6.558	3.885	3.042	1.169

Table 5 presents the results of the comparison of test data, where the SVM algorithm takes a test time of 1.402 seconds to 0.198 with a data ratio of 80:20. Then at the data ratio of 70:30, it changes from 0.921 seconds to 0.291, while the 60:40 showed a time changed from 1.241 seconds to 0.552. Also, the algorithm for each data ratio takes 0.841 to 0.698, 0.718 to 0.108, and 0.115 to 0.166. Based on these results, both algorithms tend to decrease or have a more efficient test time.

Table 5. Computational time in testing process of random sample (in seconds)

Data Ratio (Train:Test)	All Features (s)		Forward Selection	
	SVM	RF	SVM	RF
80:20	1.402	0.841	0.198	0.698
70:30	0.921	0.718	0.291	0.108
60:40	1.241	0.115	0.552	0.166

3.2. Test results with cross-validation

According to Table 6, the comparison of test results using 3-fold cross-validation shows that the accuracy value in the forward selection is more significant than the previous one. Also, the time column of training and testing decreased after the selection because the reduction in the number of features is directly proportional to the time needed for the computing process. According to Table 7, the 3-folds validation of SVM has a difference of 1.227 seconds, while RF has 1.915. However, at 10-folds, the difference is 6.008 seconds for SVM and 3.511 for the RF algorithm. Hence, the comparison of test time in Table 6 is relatively decreased between before and after feature selection. Table 8 also showed that the 3-folds validation of SVM has a difference of 0.24 seconds, while RF has 0.149. However, for the 10-folds, the difference is 0.459 seconds for SVM and 0.467 for the RF algorithm. This finding also indicates the comparison of test time in Table 8 also decreases between before and after feature selection.

Table 6. Accuracy results of the cross-validation test

k-fold	All Features		Forward Selection	
	SVM	RF	SVM	RF
3-fold	97.1%	93.8%	97.7%	94.8%
10-fold	97.1%	94.1%	97.7%	94.9%

Table 7. Computational time in cross-validation test (in seconds)

k-fold	All Features		Forward Selection	
	SVM	RF	SVM	RF
3-fold	2.115	2.143	0.888	0.228
10-fold	9.989	4.383	3.981	0.872

As seen in the computational process and the accuracy value, the overall result obtained before and after the feature selection significantly changed. Forward feature selection generally impacts training time, testing, and accuracy results. However, this finding shows that feature selection does not guarantee an increase in accuracy or computation time improvements. Therefore, more tests are needed.

3.3. Confusion matrix (CM) evaluation with cross-validation

After conducting the test using cross-validation and random sample results, the next step is to process the CM to examine the number of instances that were not predicted. Table 8 also shows that less than 5% of instances were not predicted, indicating 1.900 instances are successfully predicted. Therefore, the difference before and after being predicted is 1%. However, it is 4.5% true negative instances when using the RF algorithm (an increase in 80 instances was successfully predicted).

The results of SVM algorithms between Tables 8 and 9 do not differ significantly since the 3-fold and 10-fold have the same predicted instances results. Whereas in the RF algorithm, the difference is only at 0.01%, which means that the number of folds does not increase the accuracy of the prediction. Moreover, the terms "0" and "1" in Table 8, Table 9, and Table 10 mean the class of toxic and non-toxic, respectively.

Table 8. Confusion matrix of 3-fold cross-validation with 2,000 instances

Classification		All Features		Forward Selection	
		0	1	0	1
SVM	0	99.2%	4.6%	99.2%	3.6%
	1	0.1%	95.4%	0.15%	96.4%
RF	0	99.2%	9.1%	94.1%	4.8%
	1	0.8%	90.9%	5.9%	95.2%

Table 9. Confusion matrix of 10-fold cross-validation with 2,000 instances

Classification		All Features		Forward Selection	
		0	1	0	1
SVM	0	99.2%	4.6%	99.2%	3.6%
	1	0.0%	95.4%	0.0%	96.4%
RF	0	98.9%	9.2%	94.2%	4.7%
	1	1.1%	90.8%	5.8%	95.3%

Table 10. Confusion matrix of 3 data ratios

Classification		All Features		Forward Selection	
		0	1	0	1
		80:20 Ratio			
SVM	0	99.2%	4.6%	99.2%	3.6%
	1	0.0%	95.4%	0.0%	96.4%
RF	0	98.9%	9.2%	94.2%	4.7%
	1	1.1%	90.8%	5.8%	95.3%
		70:30 Ratio			
SVM	0	99.2%	5.0%	99.2%	4.2%
	1	0.0%	95.0%	0.0%	95.8%
RF	0	99.5%	9.0%	94.6%	5.0%
	1	0.50%	91.0%	5.4%	95.0%
		60:40 Ratio			
SVM	0	99.2%	5.0%	99.2%	4.0%
	1	0.0%	95.0%	0.0%	96.0%
RF	0	99.3%	9.10%	95.4%	5.70%
	1	0.70%	90.9%	4.6%	94.3%

3.4. Confusion matrix (CM) evaluation with random sample

Table 10 compares the results of instance predictions with ratios of three different data. In the 80:20 with the SVM algorithm, the prediction of true negative instances increased by 1.2%. The RF algorithm has increased much higher to 4.2%, from the initial value of 91% to 95.2%. Furthermore, in the 70:30 data ratio, SVM increased by 0.8% true negative variable, while RF escalated by 4%. At a ratio of 60:40, SVM also intensified by 1%, and RF increased by 3.4% on true negative variables. Based on these results, the improvement in the SVM prediction rate for each data ratio did not experience any decrease in the level of prediction accuracy but persisted at 99.2%. Therefore, the SVM algorithm tends to enhance accuracy but not as high as RF because its difference from RF is only 0.8-1.4% with a 4% rise. Moreover, Table 11 shows a comparison between our proposed work and previous work. Our proposed feature selection work improves the accuracy into 99.5%, the number which is considered a common target of Deep Learning researchers.

Table 11. Comparison of work

Author	Year	Deep Learning Model	Accuracy
Sharma *et al*. [42]	2018	Naïve Bayes, SVM, and RF	73.42% (using Naïve Bayes), 71.71% (using SVM), and 76.42% (using RF)
Oriola and Kotzé [43]	2019	SVM	95.74%
Juuti *et al*. [44]	2020	Generative Pre-trained Transformer 2 (GPT-2)	97.3%
d'Sa *et al*. [45]	2020	BiLSTM-CNN	97%
Malik *et al*. [7]	2021	BiLSTM and CNN	96.2% (using BiLSTM) and 95.42% (using CNN)
Our Proposed Work	2021	SVM and RF	99.2% (using SVM) and 99.5% (using RF)

4. CONCLUSION

Toxic speech is still a continual threat; hence every piece of information is implicitly dangerous, especially when it is not prevented. Consequently, a classification of toxic speech using the SVM algorithm with the cross-validation test method and random samples obtained is considered appropriate to solve this problem. Our work's accuracy, training time, and test results have shown a positive trend, compared to the original feature's results. A total of 10 features were obtained after a selection process from 72 initial ones. The wrapper method with the forward feature selection technique also increased the computation time and accuracy by 90%, which increases up to 5% in the RF algorithm and 0.4%-1.2% for the SVM algorithm. Thus, the forward selection is suitable for classifying toxic speech features. Also, it is implied that the number of features in the classification does not guarantee predictions with a high degree of accuracy. We expect some future work, such as the employment of different speech features of mel-frequency cepstral coefficient (MFCC), Voice Onset Time, Signal Noise-to-Ratio (SNR), speech rate, and many more.

ACKNOWLEDGEMENTS

We thank the Kementerian Riset dan Teknologi (Indonesian Ministry of Research and Technology) and the Badan Riset dan Inovasi Nasional (Indonesian National Research and Innovation Agency) for providing the research grant with contract number 002/AMD-SP2H/LT-MULTI-PDPK/LPPM-UKP/2021 (under the common contract number 001/AMD-SP2H/LT-MULTI-PDPK/LL7/2021).

REFERENCES

[1] L. Tirrell, "Toxic Speech: Inoculations and Antidotes," *South. J. Philos.*, vol. 56, pp. 116–144, 2018, doi: 10.1111/sjp.12297.

[2] A. Koratana and K. Hu, "Toxic Speech Detection," *Neural Inf. Process. Syst.*, p. 9, 2018.

[3] J. Risch and R. Krestel, "Toxic Comment Detection in Online Discussions," in *Deep Learning-Based Approaches for Sentiment Analysis*, Springer, 2020, pp. 85–109.

[4] N. I. Pratiwi, I. Budi, and M. A. Jiwanggi, "Hate Speech Identification using the Hate Codes for Indonesian Tweets," in *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, 2019, pp. 128–133.

[5] H. Mohaouchane, A. Mourhir, and N. S. Nikolov, "Detecting Offensive Language on Arabic Social Media using Deep Learning," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 466–471.

[6] M. Williams, M. Butler, A. Jurek-Loughrey, and S. Sezer, "Offensive Communications: Exploring the Challenges Involved in Policing Social Media," *Contemp. Soc. Sci.*, vol. 16, no. 2, pp. 227–240, Mar. 2021, doi: 10.1080/21582041.2018.1563305.

[7] P. Malik, A. Aggrawal, and D. K. Vishwakarma, "Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 1254–1259, doi: 10.1109/ICCMC51019.2021.9418395.

[8] S. Mishra, S. Prasad, and S. Mishra, "Exploring Multi-Task Multi-Lingual Learning of Transformer Models for Hate Speech and Offensive Speech Identification in Social Media," *SN Comput. Sci.*, vol. 2, no. 2, p. 72, Apr. 2021, doi: 10.1007/s42979-021-00455-5.

[9] S. T. Roweis, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science (80-.).*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.

[10] J. Rong, G. Li, and Y. P. P. Chen, "Acoustic Feature Selection For Automatic Emotion Recognition from Speech," *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, 2009, doi: 10.1016/j.ipm.2008.09.003.

[11] V. Bombardier and L. Wendling, "Multi-Scale Fuzzy Feature Selection Method applied to Wood Singularity Identification," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 1, p. 108, 2018, doi: 10.2991/ijcis.2018.25905185.

[12] R. Kamala and R. J. Thangaiah, “An Improved Hybrid Feature Selection Method for Huge Dimensional Datasets,” *IAES Int. J. Artif. Intell.*, vol. 8, no. 1, p. 77, Mar. 2019, doi: 10.11591/ijai.v8.i1.pp77-86.

[13] A. G. Sooai, Khamid, K. Yoshimoto, H. Takahashi, S. Sumpeno, and M. H. Purnomo, “Dynamic Hand Gesture Recognition on 3D Virtual Cultural Heritage Ancient Collection Objects using k-Nearest Neighbor,” *Eng. Lett.*, vol. 26, no. 3, pp. 356–363, 2018.

[14] A. G. Sooai, “Comparison of Recognition Accuracy on Dynamic Hand Gesture using Feature Selection,” *2018 Int. Conf. Comput. Eng. Netw. Intell. Multimed.*, pp. 270–274.

[15] A. Loukina, K. Zechner, L. Chen, and M. Heilman, “Feature Selection for Automated Speech Scoring,” in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 12–19, doi: 10.3115/v1/w15-0602.

[16] D. Kristomo, R. Hidayat, I. Soesanti, and A. Kusjani, “Heart Sound Feature Extraction and Classification using Autoregressive Power Spectral Density (AR-PSD) and Statistics Features,” in *AIP Conference Proceedings*, 2016, p. 090007, doi: 10.1063/1.4958525.

[17] M. K. Hasan *et al.*, “UR-FUNNY: A Multimodal Language Dataset for Understanding Humor,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2046–2056, doi: 10.18653/v1/D19-1211.

[18] G. K. Birajdar and M. D. Patil, “Speech/Music Classification using Visual and Spectral Chromagram Features,” *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 1, pp. 329–347, Jan. 2020, doi: 10.1007/s12652-019-01303-4.

[19] T. Drugman, M. Gurban, and J. P. Thiran, “Relevant Feature Selection for Audio-Visual Speech Recognition,” *2007 IEEE 9Th Int. Work. Multimed. Signal Process. MMSP 2007 - Proc.*, pp. 179–182, 2007, doi: 10.1109/MMSP.2007.4412847.

[20] M. Bezoui, “Speech Recognition of Moroccan Dialect using Hidden Markov Models,” *IAES Int. J. Artif. Intell.*, vol. 8, no. 1, p. 7, Mar. 2019, doi: 10.11591/ijai.v8.i1.pp7-13.

[21] X. Ying, “An Overview of Overfitting and its Solutions,” in *Journal of Physics: Conference Series*, 2019, vol. 1168, no. 2, p. 22022, doi: 10.1088/1742-6596/1168/2/022022.

[22] D. Ververidis and C. Kotropoulos, “Fast and Accurate Sequential Floating Forward Feature Selection with the Bayes Classifier applied to Speech Emotion Recognition,” *Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008, doi: 10.1016/j.sigpro.2008.07.001.

[23] B. Y. A. Schofield and T. Davidson, “Identifying Hate Speech in Social Media,” *XRDS Crossroads, ACM Mag. Students*, vol. 24, no. 2, 2017.

[24] F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE - The Munich versatile and fast open-source audio feature extractor,” *MM’10 - Proc. ACM Multimed. 2010 Int. Conf.*, no. January 2010, pp. 1459–1462, 2010, doi: 10.1145/1873951.1874246.

[25] A. Koduru, H. B. Valiveti, and A. K. Budati, “Feature Extraction Algorithms to Improve the Speech Emotion Recognition Rate,” *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, Mar. 2020, doi: 10.1007/s10772-020-09672-4.

[26] B. K. Baniya, D. Ghimire, and J. Lee, “Automatic Music Genre Classification using Timbral Texture and Rhythmic Content Features,” *Int. Conf. Adv. Commun. Technol.*, no. 3, pp. 434–443, 2015, doi: 10.1109/ICACT.2015.7224907.

[27] D. N. S. Handayani and Y. Pramudya, “Analysis of Sound Frequency and Sound Intensity in the Cylindrical Musical Instrument using Audacity Software,” *Proceeding of ICMSE*, vol. 4, no. 1, pp. 171–176, 2017.

[28] C. Gorris *et al.*, “Acoustic Analysis of Normal Voice Patterns in Italian Adults by using Praat,” *J. Voice*, 2019, doi: 10.1016/j.jvoice.2019.04.016.

[29] J. P. Teixeira, C. Oliveira, and C. Lopes, “Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters,” *Procedia Technol.*, vol. 9, pp. 1112–1122, 2013, doi: 10.1016/j.protcy.2013.12.124.

[30] P. Boersma, “Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound,” *Proc. Inst. Phonetic Sci.*, vol. 17, pp. 97–110, 1993.

[31] H. Aouani and Y. Ben Ayed, “Speech Emotion Recognition with Deep Learning,” *Procedia Comput. Sci.*, vol. 176, pp. 251–260, 2020, doi: 10.1016/j.procs.2020.08.027.

[32] B. Nouri-Moghaddam, M. Ghazanfari, and M. Fathian, “A Novel Multi-Objective Forest Optimization Algorithm for Wrapper Feature Selection,” *Expert Syst. Appl.*, vol. 175, p. 114737, Aug. 2021, doi: 10.1016/j.eswa.2021.114737.

[33] V. Kadam, “Thematic Issue Intelligent Data Mining for Data Analysis and Knowledge Discovery,” *Recent Adv. Comput. Sci. Commun.*, vol. 13, no. 3, pp. 433–434, Aug. 2020, doi: 10.2174/266625581303200609114251.

[34] G. Dong and H. Liu, *Feature Engineering for Machine Learning and Data Analytics*. CRC Press, 2018.

[35] O. Pierre-Yves, “The Production and Recognition of Emotions in Speech: Features and Algorithms,” *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1–2, pp. 157–183, Jul. 2003, doi: 10.1016/S1071-5819(02)00141-6.

[36] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: applications, challenges and trends,” *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[37] F. Tscherne, N. Wilke, B. Schachenhofer, K. Roux, and G. Tavlaridis, “Orange: Data Mining Toolbox in Python,” *Int. J. Conserv. Sci.*, vol. 7, no. SpecialIssue1, pp. 295–300, 2016.

[38] P. Probst, M. N. Wright, and A. Boulesteix, “Hyperparameters and Tuning Strategies for Random Forest,” *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3, May 2019, doi: 10.1002/widm.1301.

[39] D. P. Adi, L. Junaedi, Frismanda, A. B. Gumelar, and A. A. Kristanto, “Exploring the Time-efficient Evolutionary-based Feature Selection Algorithms for Speech Data under Stressful Work Condition,” *Emit. Int. J. Eng. Technol.*, vol. 9, no. 1, pp. 60–74, Feb. 2021, doi: 10.24003/emitter.v9i1.571.

[40] F. Khan, S. Kanwal, S. Alamri, and B. Mumtaz, “Hyper-parameter Optimization of Classifiers, using an Artificial Immune Network and its Application to Software Bug Prediction,” *IEEE Access*, vol. 8, pp. 20954–20964, 2020.

[41] A. Luque, A. Carrasco, A. Mart\'\in, and A. de las Heras, “The Impact of Class Imbalance in Classification Performance Metrics based on the Binary Confusion Matrix,” *Pattern Recognit.*, vol. 91, pp. 216–231, 2019.

[42] S. Sharma, S. Agrawal, and M. Shrivastava, “Degree-based Classification of Harmful Speech using Twitter Data,” *arXiv Prepr. arXiv1806.04197*, 2018.

[43] O. Oriola and E. Kotzé, “Automatic Detection of Toxic South African Tweets Using Support Vector Machines with N-Gram Features,” in *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 2019, pp. 126–130.

[44] M. Juuti, T. Gröndahl, A. Flanagan, and N. Asokan, “A Little goes a Long Way: Improving Toxic Language Classification despite Data Scarcity,” *arXiv Prepr. arXiv2009.12344*, 2020.

[45] A. G. d’Sa, I. Illina, and D. Fohr, “BERT and fastTEXT Embeddings for Automatic Detection of Toxic Speech,” in *2020 International Multi-Conference on:“Organization of Knowledge and Advanced Technologies”(OCTA)*, 2020, pp. 1–5.

BIOGRAPHIES OF AUTHORS

Agustinus Bimo Gumelar received a Bachelor's degree in Industrial Engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2004, and the Master Degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2010 with Honour. He is currently working toward the Ph.D degree in Electrical Engineering at Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, since 2019. His research interests include auditory-based neuroscience and affective computing, signal processing, and computational intelligence. He can be contacted at email: bimogumelar@ieee.org.

Astri Yogatama received a Bachelor's degree from the University of Brawijaya Malang in 2001 and a Master's degree in Communication Science from Airlangga University in 2009. Currently, she is the head of the Public Relation Laboratory in the Faculty of Communication Science of Petra Christian University. Her research interests include media communication, corporate communication, and strategic communication. She can be contacted at email: astri@petra.ac.id.

Derry Pramono Adi has joined the Department of Electrical Engineering, at Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, since late 2021, in pursuit of his Master's Degree. He received his Bachelor's Degree in Computer Science from Universitas Narotama recently in early 2021. Currently, he is also the Student Member of the IEEE. His current primary field of interest is affective computing, social sciences, machine learning, and natural language processing. He can be contacted at email: derryalbertus@ieee.org.

Frismanda is a last-year Computer Science student at the Universitas Narotama, Surabaya, Indonesia. He joined the undergraduate program at Universitas Narotama in 2016. His primary field of interest currently lies in machine learning. He can be contacted at email: frismanda@gmail.com.

Indar Sugiarto is the Associate Professor of Petra Christian University, Surabaya, Indonesia. He did his post-doctoral research at the University of Manchester from 2015 to 2018. He received the Ph.D. degree from Technische Universität München in 2015, and M.Sc. degree from Universität Bremen. His current research interests include artificial intelligence, neuromorphic computing, and robotics. He can be contacted at email: indi@petra.ac.id.