Optimizing Random Forest for East Java Election Sentiment with Chi-Square and Mutual Information

Rahma Putri Widyaiswari[1], Anisa Dzulkarnain[2*], Alqis Rausanfita[3]
[1,2] *Department of Information System, Telkom University, Surabaya Campus, Indonesia*
[3] *Department of Informatics, Telkom University, Surabaya Campus, Indonesia*
*corr-author: anisadzulkarnain@telkomuniversity.ac.id

Abstract - The rise of social media has transformed the way people express opinions, including in political contexts. In the 2024 East Java Gubernatorial Election, social media platform X became a major outlet for public sentiment toward the governor and deputy governor candidates. This study aims to analyse public sentiment toward three candidate pairs by categorizing the data into three sentiment classes: positive, negative, and neutral. Feature selection was conducted by combining Term Frequency-Inverse Document Frequency (TF-IDF) with Chi-Square and Mutual Information (MI) methods to improve feature quality. The Random Forest algorithm was employed as the primary classification model. In addition, several other algorithms were tested for comparison. The results indicate that the TF-IDF and Chi-Square combination with Random Forest achieved the highest accuracy of 82.07%. These findings highlight the importance of feature selection in improving model performance for sentiment classification. The study provides insights into public opinion that can serve as a reference for strategic decision-making in the political and public sectors.

Keywords: sentiment analysis, East Java governor election, social media X, random forest, feature selection

I. INTRODUCTION

Social media X (formerly Twitter) has developed into one of the most effective channels of political communication, especially because of its ability to convey messages quickly and reach a wide audience, including in the context of the Regional Head Election campaign [1]. Apart from being a communication medium, this platform is also a public discussion space where the public can express their views, aspirations, and criticisms of the candidates for governor and deputy governor [2]. Ahead of the 2024 East Java Regional Elections, public participation in voicing political opinions through social media has increased significantly. The various uploads that appear form a large data set that is unstructured, including the expression of positive, negative, and neutral sentiments that represent the public's perception of each candidate, while potentially influencing voter preferences [3].

In understanding the dynamics of public opinion systematically, a text mining-based sentiment analysis approach is an effective solution. Sentiment analysis allows for the evaluation of individual attitudes and opinions based on text data, such as comments on social media [4]. The Random Forest algorithm has been widely used as a reliable classification method in sentiment analysis due to its ability to handle large data with good classification performance. Previous studies have shown that Random Forest is able to achieve up to 94% accuracy on the analysis of presidential election sentiment, surpassing the Lexicon-Based method [5].

However, the performance of Random Forest can be further improved through the application of proper feature selection techniques. Feature selection aims to select the subset of features that are most relevant to the classification target and reduce the complexity of the data. Among the various methods available, Chi-Square and Mutual Information are two commonly used techniques. Chi-Square evaluates feature relevance based on the statistical relationship between features and target classes, while Mutual Information measures the level of shared information between features and classification labels [6,7]. Previous research has shown that the implementation of feature selection can significantly improve model performance. For example, the combination of Chi-Square with the Modified Balanced Random Forest algorithm was able to achieve an accuracy of 81.75% and an F1-score of 71.90% on the sentiment analysis of beauty product reviews [8]. In addition, Mutual Information has also been shown to be effective in selecting relevant features to improve classification performance [9].

Based on this background, this study aims to improve the performance of the Random Forest algorithm in analysis the sentiment of the 2024 East Java Regional Election on social media X through the application of Chi-Square and Mutual Information feature selection techniques. Random Forest was chosen as the primary

classification method because of its ability to handle large and complex data with stable performance. By integrating feature selection techniques, it is hoped that the results of sentiment classification will be more accurate and relevant, as well as make a practical contribution in supporting data-based political communication strategies that are responsive to public opinion.

II. METHOD

The methodology in this study is carried out using several stages of the process. The stages of this research are shown in Fig. 1.

The study involved collecting data from Social Media X about the 2024 East Java Election, followed by data cleaning, sentiment labeling, and preprocessing text steps. The data was then split, vectorized using TF-IDF, and selected using Chi-Square and Mutual Information. Classification was performed using Random Forest, and the model was evaluated and visualized.

A. Data Collection

The data collection in this study was obtained from social media X in the form of X posts about the 2024 East Java Provincial Elections, especially related to gubernatorial and deputy governor candidates, using keywords, which include the names of candidate pairs, supporting parties, and taglines of each pair. The data was taken using the Tweet Harvest method through Google Collab, using API (Application Programming Interface) authentication and filtering Indonesian-language X posts, with a time range of July 1 to November 28, 2024. Data collection was carried out during November 2024 and continued in April 2025 which produced around 6,445 data. The data is collected along with additional attributes such as post time, username, number of likes, and stored in CSV (Comma-Separated Values) format.

B. Data Cleaning

The data preprocessing stage was conducted to prepare the raw dataset for modelling. This study initially collected 6,445 rows of X posts data related to the 2024 East Java gubernatorial election. Several cleaning steps were performed, including the removal of irrelevant columns to maintain analytical focus, handling missing values by deleting rows or columns with excessive null entries, and eliminating duplicate records to ensure data quality and accuracy [10]. After completing all preprocessing stages, the dataset was reduced to 5,950 rows. The cleaned data was then stored in CSV format using the to_csv() function from the pandas library for further modelling.

C. Sentiment Labelling

Sentiment labelling is a process in data analysis that aims to classify reviews into positive, negative, or neutral categories in a dataset [11]. This process can be done manually or automatically. In this study, sentiment labelling is carried out manually using three classes: positive, negative, and neutral. After the data is prepared, this labelling stage is conducted to assign each review to the appropriate sentiment category. To ensure consistency and reduce bias, the labelling process involved three annotators guided by a predefined annotation guideline. One annotator performed the main labelling task, while two additional annotators verified a random subset of the data. In case of disagreement, the final label was decided through discussion based on the agreed criteria. The results of this process are presented in Table I.

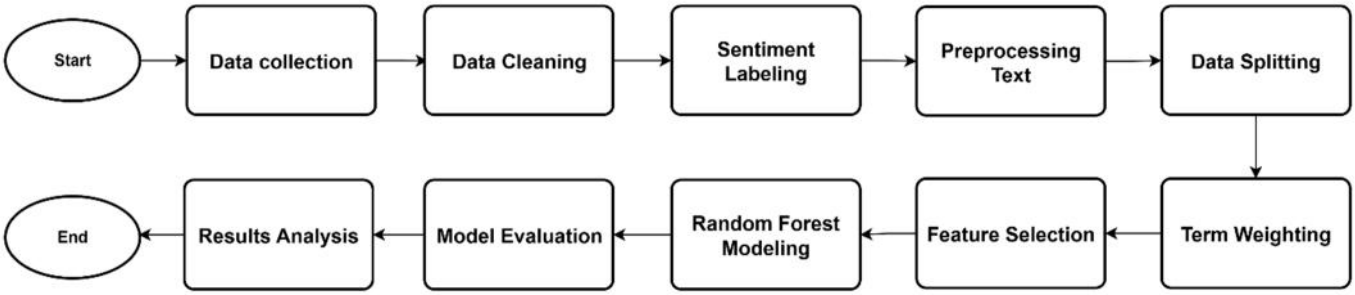

Fig. 1 Research method flow

TABLE I
SENTIMENT LABELLING

Text	Label
@aniesbaswedan @lulukhamidah @DPP_PKB @cakimiNOW Keren banget bisa dapet kesempatan diskusi sama mbak Luluk Hamidah. Semoga bisa nambah wawasan dan semangat kita! (@aniesbaswedan @lulukhamidah @DPP_PKB @cakimiNOW It's amazing to have the opportunity to have a discussion with Ms. Luluk Hamidah. Hopefully, it can broaden our insights and boost our enthusiasm!)	1
@NenkMonica Tenggelam kan Khofifah Indar parawansa dan Emil Dardak (@NenkMonica Let Khofifah Indar Parawansa and Emil Dardak sink!)	2
@luuutpi Aku bingung mau milih bu risma bu khofifah apa bu luluk yahhh? (@luuutpi I'm confused should I choose Mrs. Risma, Mrs. Khofifah, or Mrs. Luluk?)	0

Based on the results of sentiment labelling, the overall distribution of data can be seen in Fig. 2. The graph shows that positive sentiment dominates with 2431 data, followed by neutral sentiment with 2103 data, and negative sentiment with 1404 data. In addition, the distribution of sentiment for each candidate is shown in Fig. 2 which shows the proportion of positive, neutral, and negative sentiments received by each candidate based on the data that has been labelled.

In addition to the overall distribution of sentiment, Fig. 3 also shows the detailed distribution of sentiment for each candidate to illustrate the proportion of positive, neutral, and negative opinions for each candidate pair.

Fig. 3 presents the sentiment analysis results for each candidate. Luluk–Lukman received the highest positive sentiment (42.4%), followed by neutral (36.3%) and negative (21.3%), indicating generally favourable public opinion. Khofifah was also dominated by positive sentiment (40.5%), but had the highest negative sentiment (27.9%), suggesting mixed perceptions with notable criticism. Risma showed the highest overall positivity (40.3%) and lowest negativity (20.1%), with a narrow gap between positive and neutral (39.7%), reflecting a strong yet balanced public image.

D. Preprocessing Text

The preprocessing phase of Table II text involves a series of structured steps to clean and standardize textual data prior to modelling. It starts with cleanup, which removes unnecessary elements such as punctuation, symbols, URLs, emoticons, and hashtags using a re library [11]. Uppercase folding is then applied by converting all characters to lowercase for consistency[11,12]. This is followed by normalization to correct informal words and spellings based on predetermined dictionaries [13]. The cleaned text was then tokenized using NLTK, followed by the removal of the break words using a combination of NLTK, literature-based, and a manually curated list of end words [14]. Finally, stemming is done using the Literary library to reduce words to their root form [15].

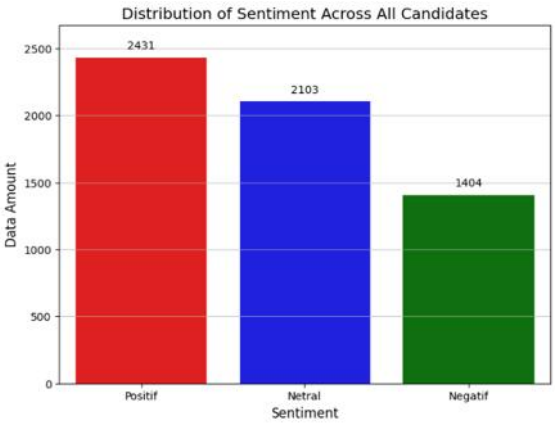

Fig. 2 Distribution of sentiment across all candidates

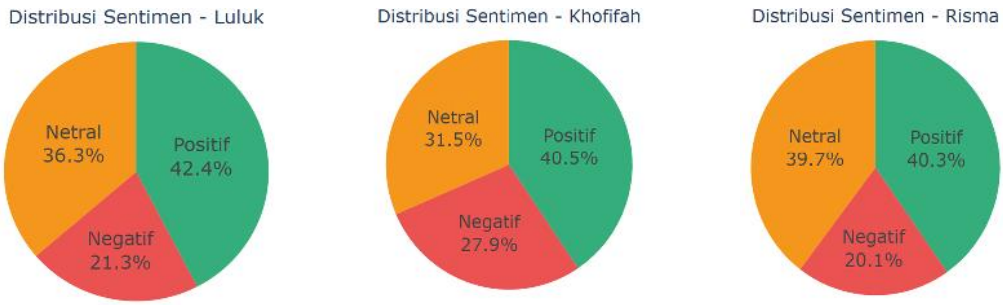

Fig. 3 Distribution of sentiment labels for each candidate in the 2024 east java gubernatorial election

TABLE II
TEXT PREPROCESSING

Process	Results
Initial Text	*@OmahJTM Jatim tuh butuh perubahan yang nyata bukan cuma omongan kosong. Khofifah dan Emil udah buktiin dengan kerja keras mereka* (@OmahJTM East Java needs real change, not just empty words. Khofifah and Emil have already proven it through their hard work)
Cleaning	*Jatim tuh butuh perubahan yang nyata bukan cuma omongan kosong Khofifah dan Emil udah buktiin dengan kerja keras mereka* (East Java needs real change, not just empty promises Khofifah and Emil have already proven themselves through their hard work)
Case Folding	*jatim tuh butuh perubahan yang nyata bukan cuma omongan kosong khofifah dan emil udah buktiin dengan kerja keras mereka* (east java needs real change, not just empty promises khofifah and emil have proven it through their hard work)
Normalization	*jawa timur itu butuh perubahan yang nyata bukan cuma omongan kosong khofifah dan emil sudah buktikan dengan kerja keras mereka* (east java needs real change, not just empty words khofifah and emil have proven it through their hard work)
Tokenization	*['jawa', 'timur', 'itu', 'butuh', 'perubahan', 'yang', 'nyata', 'bukan', 'cuma', 'omongan', 'kosong', 'khofifah', 'dan', 'emil', 'sudah', 'buktikan', 'dengan', 'kerja', 'keras', 'mereka']* (['Java', 'East', 'that', 'need', 'change', 'which', 'real', 'not', 'just', 'talk', 'empty', 'khofifah', 'and', 'emil', 'already', 'prove', 'with', 'work', 'hard', 'they']
Stopword Removal	*['jawa', 'timur', 'butuh', 'perubahan', 'nyata', 'cuma', 'omongan', 'kosong', 'khofifah', 'emil', 'buktikan', 'kerja', 'keras']* (['java', 'east', 'need', 'change', 'real', 'only', 'talk', 'empty', 'khofifah', 'emil', 'prove', 'work', 'hard'])
Stemming	*jawa timur butuh ubah nyata cuma omong kosong khofifah emil bukti kerja keras* (east java needs real change, not just empty words khofifah and emil should show proof of hard work)

E. Splitting Data

After preprocessing, the dataset is split into training and testing sets to evaluate model performance and prevent overfitting [16]. A total of 5,938 preprocessed X posts are randomly divided into three data split ratios: 90:10, 80:20, and 70:30. This ensures a proportional sentiment distribution across both sets, allowing the model to train effectively and be tested on unseen data. Table III presents the number of samples in each split.

F. Term Weighting

In this stage, the TF-IDF method is used to determine the importance of a term within a document. Term Frequency (TF) measures how often a term appears in a single document, while Inverse Document Frequency (IDF) reflects how rare the term is across the entire corpus. A higher TF-IDF score indicates that the term is more relevant to the specific document [17]. The calculation in this study follows the formula shown in equation (1).

$$W(t,d) = TF(t,d) \cdot IDF(t) \quad (1)$$

TF-IDF weighting was then applied using scikit-learn's TfidfVectorizer, converting texts into numerical values that reflect term importance. The results are shown in Table IV.

Table V shows the TF, DF, and TF-IDF values for five sample words. The word "micro" has the highest TF-IDF score (0.405972), indicating high importance due to its limited but frequent presence. In contrast, "risma" has the lowest score (0.054359), reflecting its common use across many documents and lower discriminative power. This illustrates how TF-IDF captures both frequency and word specificity.

G. Selection Features

In this study, the feature selection process aims to improve the accuracy of the classification model and reduce the dimension of irrelevant data. The two methods used are MI (Mutual Information) and Chi-Square.

TABLE III
DISTRIBUTION OF TRAINING AND TESTING DATA BY SPLIT RATIO

Ratio Split	Training Data	Testing Data	Total Data
90:10	5.344	594	5.938
80:20	4.750	1.188	5.938
70:30	4.157	1.781	5.938

TABLE IIV
WORD WEIGHTING

Word	TF	DF	TF-IDF
tanding (match)	1	8	7.386692
telak (crushing)	1	32	6.087409
dapat (can)	1	73	5.279852
menang (win)	1	492	3.383408
emil	1	997	2.678164
khofifah	1	2498	1.760271

1) MI (Mutual Information): Use it to measure how much reliance a feature has on the target label. The higher the MI score, the greater the contribution of the feature to the class information. The formula MI refers to (2) [9].

$$I(U,C) = \sum_{et\in\{1,0\},ec\in\{1,0\}} P(U=et, C=ec)\log_2 \frac{P(U=et, C=ec)}{P(U=et)?P(C=ec)} \quad (2)$$

In (2) it is theoretical because it states the relationship between the joint probability and the marginal probability of two binary variables, without relying directly on empirical data. The main objective is to measure the degree of dependence between the two variables; the greater the MI value, the more information the two variables provide each other. This principle is then applied in the feature selection process, where the TF-IDF method is combined with Mutual Information (MI) to select the words with the highest relevance to the classification target. Some selected words are shown in Table V as an important feature in the sentiment classification model.

Based on Table V, the word "khofifah" has the highest Mutual Information score, followed by the words "emil" and "*menang* (win)". This shows that these words have the highest relevance in helping the model distinguish sentiment classes. Overall, out of the 6563 initial features, 4594 features (70% with the highest scores) were selected through a feature selection process using Mutual Information.

2) Chi-Square: to evaluate the statistical relationship between features and classes in the form of an independence test. A high Chi-Square value indicates that the feature has a significant relationship with the target class. Chi-Square calculation formula refers to equation (3) [18]:

$$i^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

In equation (3), where (i^2) used to evaluate the degree of difference between observation values O_{ij} and expectation values E_{ij} on each cell in the contingency table. The calculation adds up all the results of the division between the square of the difference between the observation value and the expected value to the expected value at the position of the cell (i, j). This statistical principle is applied in the feature selection process, where the Chi-Square (X^2) method is combined with TF-IDF to select the words most relevant to the classification target. Table VI presents the selected words with the highest Chi-Square scores.

Table VI shows the word "emil" having the highest score (23.144108), followed by "*menang* (win)" and "*dapat* (can)". All the words in the table are chosen as an important feature because they have the highest score based on the Chi-Square method.

H. *Random Forest Algorithm Classification Modelling*

Random Forest is a method that can improve the accuracy of classification through the process of selecting random attributes for each node. The tree formation process in Random Forest is similar to the Classification and Regression Tree (CART) method, but *Random Forest* does not prune [19]. In equation (4), *Gini* is the value (impurity) of a node that is used to measure how pure the data in a classification node is.

$$Gini = 1\sum_i^C (°_i)^2 \quad (4)$$

TABLE V
FEATURE SELECTION USING MUTUAL INFORMATION

Word	Mutual Information Score	Selected
khofifah	0.513768	True
emil	0.189959	True
menang (win)	0.095918	True
dapat (can)	0.014164	True
telak (crushing)	0.006204	True
tanding (match)	0.001570	True

TABLE VI
FEATURE SELECTION USING CHI-SQUARE

Word	Chi-Square (X^2) Score	Selected
emil	23.144108	True
menang (win)	14.309978	True
dapat (can)	5.777478	True
Telak (crushing)	3.063921	True
khofifah	0.965070	True
tanding (match)	0.961808	True

Eq. (4), indicates symbol ∑ the addition operation, where *i* is an index of classes starting from 1 to *C*, With *C* as the total number of classes available. Notation *ρi* expresses the proportion of data from class *I* in a node, while *ρi²* is the square of the proportion. *Gini impurity* calculated based on the sum of the values *ρi²* for the entire class, then subtracted from 1. The smaller the value *Gini*, then the purer the node. In equation (5), *Gain*(*A*) it is a measure of the gain of information obtained from the separation of data using the *A*.

$$gain(A) = Gini\ impurity - Gini\ average \quad (5)$$

The value in (5) is calculated by subtracting *Gini impurity*, namely value *Gini* of nodes before separation, with *Gini Average*, namely the average value *Gini* of *the nodes* of the separation result by the *A*. The greater the value *Gini*, the better the attribute at dividing the data into purer classes.

I. Model Evaluation

Model evaluation using the Confusion Matrix is used to evaluate the performance of the classification model by displaying the number of true and false predictions as seen in Table VII [20]. Evaluations were conducted based on accuracy, precision, recall, and F1-score to assess the accuracy, coverage, and balance of model performance [21].

Table VII summarizes the classification outcomes across the three sentiment classes. TPP (True Positive Positive) refers to the number of positive instances correctly predicted. PFNEG and PFNET represent negative and neutral data that were misclassified as positive. For the negative class, TNEGNEG indicates correct predictions, while NEGFP and NEGFNET are positive and neutral instances mistakenly labelled as negative. Meanwhile, TNETNET shows neutral data accurately classified, with NETFP and NETFNEG representing positive and negative data wrongly predicted as neutral. These components form the basis for calculating evaluation metrics such as accuracy. Where accuracy is used to measure the accuracy of the overall prediction of the model based on the comparison between the correct prediction and the total data. The calculation of accuracy refers to (6).

$$A = \frac{(T + T. + T.)}{(T A)} \times 1 \quad (3)$$

III. RESULT AND DISCUSSION

Model evaluation was conducted using the Random Forest algorithm (n_estimators=400) with three vectorization methods: TF-IDF, TF-IDF + Mutual Information (MI), and TF-IDF + Chi-Square (Chi2). To handle class imbalance, SMOTE was also applied. Testing was done under three data split ratios: 90:10, 80:20, and 70:30. Accuracy is used to compare results across methods and configurations, as seen in Table VIII.

Table VIII shows the comparison of classification methods across different data split ratios. Among all methods tested, Random Forest combined with TF-IDF and Chi-Square achieved the highest accuracy (0.8207) at the 80:20 split. Overall, Random Forest outperformed Logistic Regression and Decision Tree across all configurations, especially when paired with balancing (SMOTE) or feature selection techniques (MI, Chi2). Logistic Regression and Decision Tree models yielded lower performance, indicating that Random Forest is more robust for this sentiment classification task.

In addition to accuracy comparisons, a visualization of the confusion matrix on the best-performing model is displayed to provide an overview of the prediction distribution for each sentiment class.

TABLE VII
CONFUSION MATRIX

Actual Class	Prediction Class		
	Positive	**Negative**	**Neutral**
Positive	TPP	PFNEG	PFNET
Negative	NEGFP	TNEGNEG	NEGFNET
Neutral	NETFP	NETFNEG	TNETNET

TABLE VIII
METHOD COMPARISON

Method	90:10	80:20	70:30
Random Forest Baseline	0.792929	0.804714	0.804714
Random Forest + MI	0.799663	0.819865	0.809764
Random Forest + Chi2	0.799663	**0.820707**	0.800786
Random Forest + SMOTE	0.819865	0.819865	0.814254
Logistic Regression Baseline	0.772727	0.779461	0.772166
Logistic Regression + MI	0.799663	0.777778	0.771044
Logistic Regression + Chi2	0.762626	0.778620	0.767116
Logistic Regression + SMOTE	0.776094	0.781145	0.773288
Decision Tree Baseline	0.698653	0.716330	0.687430
Decision Tree + MI	0.683502	0.731481	0.671156
Decision Tree + Chi2	0.685185	0.705387	0.660494
Decision Tree + SMOTE	0.693603	0.711279	0.691919

Fig. 4 shows that the Random Forest model with the combination of TF-IDF + Chi-Square can classify sentiment consistently, with the highest number of correct predictions in the positive class of 395 data, followed by the neutral class of 340 data, and the negative class of 240 data. Although there are still prediction errors, the model shows stable performance and good generalization capabilities in the test data. Overall, the application of the Chi-Square feature selection method has been shown to be effective in improving the model's ability to distinguish sentiment classes more accurately.

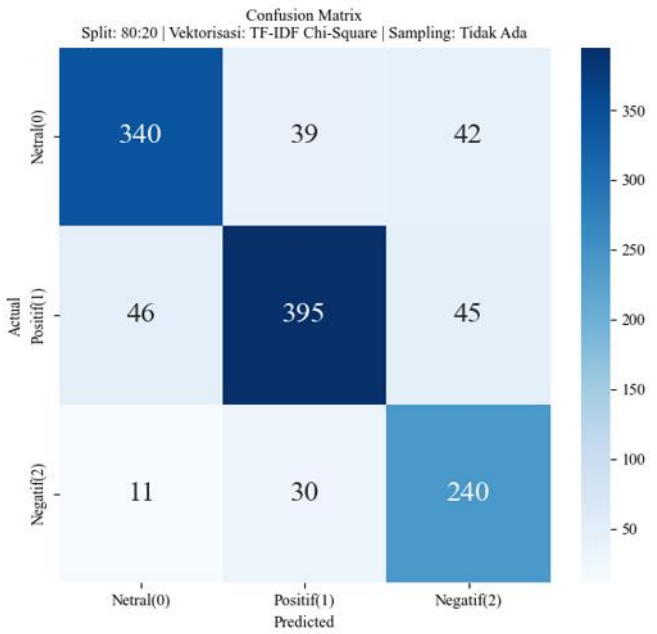

Fig. 4 Heatmap confusion matrix

IV. CONCLUSION

This study evaluates the effectiveness of the Random Forest algorithm in classifying public opinion sentiment on social media using TF-IDF features, combined with two feature selection techniques Chi-Square and Mutual Information. Experimental results show that Random Forest consistently outperforms other methods such as Logistic Regression and Decision Tree across all data split ratios (90:10, 80:20, and 70:30). The highest performance was achieved with the Random Forest + Chi-Square combination at an 80:20 split, reaching an accuracy of 82.07%, demonstrating strong classification ability and generalization. The application of feature selection techniques contributed to improving model performance over the baseline TF-IDF. In addition, models that applied sampling techniques such as SMOTE also showed improved results in several cases, particularly for Random Forest. These findings suggest that the use of Random Forest, when combined with proper preprocessing and feature selection strategies, can significantly enhance sentiment classification tasks. This research may serve as a reference for implementing similar approaches in areas such as consumer feedback analysis, public service evaluation, and political discourse monitoring.

REFERENCES

[1] T. Cahya Herdiyani and A. U. Zailani, "Sentiment Analysis Terkait Pemindahan Ibu Kota Indonesia Menggunakan Metode Random Forest Berdasarkan Tweet Warga Negara Indonesia Sentiment Analysis Related to Transportation of Indonesian Capital City Using Random Forest Method Based On Tweet Of Indonesian Citizens," *JTSI*, vol. 3, no. 2, pp. 154–165, 2022, doi: https://doi.org/10.35957/jtsi.v3i2.2920.

[2] K. Adib, M. R. Handayani, W. D. Yuniarti, and K. Umam, "Opini Publik Pasca-Pemilihan Presiden: Eksplorasi Analisis Sentimen Media Sosial X Menggunakan SVM," vol. 7, no. 2, pp. 80–91, 2024, [Online]. Available: https://doi.org/10.31598

[3] I. Arifin, M. Ahmad, J. Arifin, A. Agustang, and A. Sadriani, "Equilibrium: Jurnal Pendidikan Peran Media Sosial Dalam Mempengaruhi Keputusan Pemilih Pemula Pada Pemilu 2024 di Indonesia," vol. 12, pp. 201–208, 2024, doi: https://doi.org/10.26618/equilibrium.v12i2.14421.

[4] M. R. Fahlevvi, "Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi dan Dokumentasi Kementerian Dalam Negeri Republik Indonesia di Google Playstore Menggunakan Metode Support Vector Machine," *Jurnal Teknologi dan Komunikasi Pemerintahan*, vol. 4, no. 1, pp. 1–13, 2022, doi: https://doi.org/10.33701/jtkp.v4i1.2701.

[5] O. Manullang, C. Prianto, and N. H. Harani, "Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest," vol. 11, no. 2, pp. 159–169, 2023, doi: https://doi.org/10.33884/jif.v11i02.7987.

[6] P. W. Ratiasasadara, S. Sudarno, and T. Tarno, "ANALISIS SENTIMEN PENERAPAN PPKM PADA TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER DENGAN SELEKSI FITUR CHI-SQUARE," *Jurnal Gaussian*, vol. 11, no. 4, pp. 580–590, Feb. 2023, doi: 10.14710/j.gauss.11.4.580-590.

[7] I. Gusti, A. Ngurah, R. Semadi, M. Samsudin, and K. Dharmendra, "Perbandingan Metode Seleksi Fitur Pada Analisis Sentimen (Studi Kasus Opini PILKADA DKI 2017)," *Journal of informatics*, vol. 8, no. 1, pp. 11–18, 2023, doi: https://doi.org/10.51211/itbi.v8i1.2408.

[8] A. P. P. Wardani, A. Adiwijaya, and M. D. Purbolaksono, "Sentiment Analysis on Beauty Product Review Using Modified Balanced Random Forest Method and Chi-Square," *Journal of Information System Research (JOSH)*, vol. 4, no. 1, pp. 1–7, Oct. 2022, doi: 10.47065/josh.v4i1.2047.

[9] I. G. Putra, C. Pramartha, A. A. I. N. E. Karyawati, and M. A. Raharja, "Penerapan SVM dengan Seleksi Fitur Mutual Information untuk Memprediksi Sentimen PEMILU 2024," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 12, no. 4, pp. 2654–5101, May 2024, doi: https://doi.org/10.24843/JLK.2024.v12.i04.p11.

[10] T. Gori, A. Sunyoto, and H. Al Fatta, "Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 215–224, Feb. 2024, doi: 10.25126/jtiik.20241118074.

[11] A. Arham, E. R. Swedia, M. Cahyanti, and M. R. D. Septian, "IMPLEMENTASI SENTIMENT ANALYSIS PADA OPINI MASYARAKAT INDONESIA DI TWITTER TERHADAP VIRUS COVID-19 VARIAN OMICRON DENGAN ALGORITMA NAÏVE BAYES, DECISION TREE, DAN SUPPORT VECTOR MACHINE," *Sebatik*, vol. 26, no. 2, pp. 565–572, Dec. 2022, doi: 10.46984/sebatik.v26i2.1961.

[12] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 887–894, Aug. 2024, doi: 10.25126/jtiik.1148315.

[13] A. Zahra Latifa and W. Maharani, "Jurnal Teknik Informatika (JUTIF) Analyzing Public Sentiment on the Relocation of Indonesia's Capital to Kalimantan as the Ibu Kota Nusantara Using Logistic Regression," vol. 6, no. 2, pp. 575–592, 2025, doi: 10.52436/1.jutif.2025.6.2.4230.

[14] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," vol. 6, no. 9, pp. 4305–4313, Sep. 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[15] R. Aryanti, T. Misriati, and A. Sagiyanto, "Analisis Sentimen Aplikasi Primaku Menggunakan Algoritma Random Forest dan SMOTE untuk Mengatasi Ketidakseimbangan Data," *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 1, pp. 218–227, Nov. 2023, doi: 10.47065/josyc.v5i1.4562.

[16] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *Jurnal Sains dan Informatika*, vol. 9, no. 1, pp. 19–28, Jun. 2023, doi: 10.34128/jsi.v9i1.622.

[17] K. Hadi and E. Utami, "Analysis of K-NN with the Integration of Bag of Words, TF-IDF, and N-Grams for Hate Speech Classification on Twitter," vol. 12, no. 2, pp. 289–298, 2024, doi: https://doi.org/10.30595/juita.v12i2.23829.

[18] T. Ernayanti, M. Mustafid, A. Rusgiyono, and A. R. Hakim, "PENGGUNAAN SELEKSI FITUR CHI-SQUARE DAN ALGORITMA MULTINOMIAL NAÏVE BAYES UNTUK ANALISIS SENTIMEN PELANGGGAN TOKOPEDIA," *Jurnal Gaussian*, vol. 11, no. 4, pp. 562–571, Feb. 2023, doi: 10.14710/j.gauss.11.4.562-571.

[19] S. Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, Dec. 2022, doi: 10.35580/variansiunm31.

[20] R. Nurhidayat and K. E. Dewi, "KOMPUTA : Jurnal Ilmiah Komputer dan Informatika PENERAPAN ALGORITMA K-NEAREST NEIGHBOR DAN FITUR EKSTRAKSI N-GRAM DALAM ANALISIS SENTIMEN BERBASIS ASPEK," vol. 12, no. 1, pp.

91–100, 2023, doi: https://doi.org/10.34010/komputa.v12i1.9458.

[21] M. Andrew, A. Yasin, D. Arman Prasetya, and T. M. Fahrudin, "Analisis Sentimen Tiktok Shop Menggunakan Metode Multinomial Naïve Bayes Dan BM25," *Jurnal Ilmiah Teknologi Informasi Asia*, vol. 18, no. 02, pp. 24–31, 2024, doi: https://doi.org/10.32815/jitika.v18i2.