Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Application of Sentiment Analysis to Crime News using Tf-Idf and K-Nearest Neighbor to Assess Public Perception

Nadilla Indah Cahyani[1]*, Relita Buaton[2], I Gusti Prahmana[3]

[123]*STMIK Kaputama*
nadillaindahcahyanii@gmail.com[1*], bbcbuaton@gmail.com[2], igustiprahmana4@gmail.com[3]

Abstract

The development of information technology and social media has changed the way people access news, including criminal news that is often in the public spotlight. Criminal news not only presents facts, but can also shape public opinion quickly and widely so that it has the potential to cause disinformation. For this reason, sentiment analysis is needed that is able to provide an objective picture of public perception of criminal news.This study uses a quantitative approach with stages: collection of crime news data and public comments from online media, text *preprocessing* (cleansing, case folding, tokenizing, stopword removal, normalization, and stemming), feature extraction using *Term Frequency-Inverse Document Frequency* (TF-IDF), sentiment classification with the *K-Nearest Neighbor* algorithm (K-NN), as well as model evaluation through accuracy, precision, recall, and F1-score metrics.The results showed that the combination of TF-IDF and K-NN was able to classify public comments on criminal news into three sentiment classes (positive, negative, neutral) with an accuracy rate of 82%. Further evaluation showed an average precision value of 0.86, a recall of 0.82, and an F1-score of 0.82. These findings prove that the TF-IDF and K-NN methods are effective in understanding public perception of online media-based crime reporting.

Keywords: *Sentiment analysis, TF-IDF, K-Nearest Neighbor, criminal news, public opinion*

1. Introduction

The development of information technology and social media has brought significant changes in the way people access news, including crime news that is often in the public spotlight. Crime news not only presents the facts of events, but can also shape public opinion quickly and massively. This makes public perception often subjective, influenced by media framing and individual viewpoints, and increases the potential for disinformation. Therefore, a sentiment analysis method is needed that is able to assess public opinion more objectively and measurably.

Sentiment analysis of crime news has been proven to help understand people's views on various forms of crime that occur[1]. Emphasized that the Term Frequency-Inverse Document Frequency (TF-IDF) method is very effective in extracting text features, so that the system is able to recognize sentiment patterns more accurately. showed that the K-Nearest Neighbor (K-NN) algorithm is reliable for the classification of text data, including in determining public opinion on a criminal event.

Against this background, this study focuses on the application of the combination of TF-IDF and K-NN methods to analyze public sentiment towards criminal news. It is hoped that the results of this study can provide a more objective picture of public opinion, as well as contribute to the development of text-based sentiment analysis methods on sensitive social issues.

2. Theoretical Foundations

2.1. Sentiment Analysis

Sentiment analysis is the process of understanding, extracting, and processing textual data to obtain the sentiment information contained in it, whether it is positive, negative, or neutral. Sentiment analysis can help in understanding public opinion on certain issues by utilizing data from social media[2]. This technique has a number of advantages, including providing an understanding of the views of the community and being the basis for strategic decision-making. However, there are also limitations, such as ambiguity of language that can affect the accuracy of analysis and biased data resulting in inaccurate conclusions. In practice, sentiment analysis is widely applied in the business field to assess customer satisfaction, in the political realm to map public opinion on policies or candidates, and in social media as a means of monitoring trends and responding to public issues in real-time.

2.2. Method TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a word weighting method used to assess the importance of a word in a document compared to the overall document collection. TF-IDF is effective in extracting text features for sentiment analysis.

The main component of TF-IDF is Term Frequency (TF), which measures the frequency with which words appear in a document:

TF(t,d)=Number of words in document dNumber of occurrences of the word t in document d

$$\mathrm{TF}(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \quad (1)$$

and Inverse Document Frequency (IDF), which assesses the importance of a word in a document collection. Words that appear frequently in many documents will have a lower weight:

$$\mathrm{IDF(t)} = \log\left(1+\frac{\mathrm{N}}{\mathrm{DF(t)}}\right) \quad (2)$$

The TF-IDF value is obtained from the multiplication of the two:

$$\text{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} \quad (3)$$

With this calculation, words that rarely appear throughout the document collection but appear frequently on certain documents will have a higher weight, so they are considered more important in text analysis.

2.3. Algoritma K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) is a classification algorithm that works by searching for a number of nearby neighbor kkk from data that will be classified based on feature similarity. In sentiment analysis, K-NN is used to group text based on similarity to the trained data that has been labeled. This method can be used to analyze sentiment with a high degree of accuracy.

The K-NN work process is carried out by calculating the distance between the test data and all training data using proximity measures, one of which is Euclidean Distance, which is formulated as follows:

$$\mathrm{Distance}\left(\mathrm{x_i}\text{-}\mathrm{y_i}\right) = \sum_{i=1}^{n}\left(\mathrm{x_i}\text{-}\mathrm{y_i}\right)^2 \quad (4)$$

Information:

a) x_i and y_i is the value of the feature in the i dimension of the x and y points.
b) n is the sum of the dimensions (features) on the data.

Once the distances are calculated, the results are sorted from smallest to largest. Next, a number of the nearest neighboring kkk were selected, then the majority class of the neighbors was determined as a result of the classification of test data.

2.4. Sentiment Model Evaluation

Sentiment analysis model evaluation is performed to measure how well the model performs in classifying text into appropriate sentiment categories, such as positive, negative, or neutral. Some of the metrics that are often used in sentiment model evaluation include accuracy, precision, *recall,* and *F1-score.* The combination of these various metrics needs to be considered in order to obtain a more comprehensive picture of the model's performance, as each metric has a different measurement focus and can complement each other in providing a more thorough evaluation[3].

a. Accuracy: This metric measures how much of a prediction is correct compared to the total amount of data tested. Accuracy is calculated by the following formula:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of test data}}$$

b. Precision: Precision measures how many positive predictions are correct among all predictions that are categorized as positive. The precision formula is:

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

c. Recall: Recall measures how many positive predictions are correct among all the data that should be positive. The recall formula is:

$$\text{Recall} = \frac{\text{True Posituve}}{\text{True Positive+False Negative}}$$

d. F1-score: F1-score is the harmonic average between precision and recall, which provides an overall picture of the model's performance in terms of the balance between precision and recall. The formula of F1-score is:

$$\text{F1-Score}=2 \times \frac{\text{Presisi x recall}}{\text{Presisi+Recall}}$$

2.5. Criminal Dataset

The criminal dataset is a collection of data used as analysis material in this study. The data was obtained from criminal news taken through online media and public comments related to the news. This dataset contains text that reflects public opinion or sentiment towards a criminal event.

The use of data from social media, such as Twitter, is very useful in sentiment analysis research because it is able to provide people's views directly and in real-time. The criminal data used in this study was then processed through the preprocessing stage before feature extraction using TF-IDF and classification with the K-NN algorithm[4].

3. Analysis and Planning

3.1. Research Methods

This study uses a quantitative method with an experimental approach. The research stages include data collection in the form of crime news, text preprocessing, feature extraction with Term Frequency-Inverse Document Frequency (TF-IDF), classification using the K-Nearest Neighbor (K-NN) algorithm, and model evaluation to measure accuracy in assessing public perception.The K-NN method was chosen because it has a good ability to perform similarity-based classification between data, while TF-IDF is used to represent text into numerical form that reflects the level of importance of a word in a document collection.

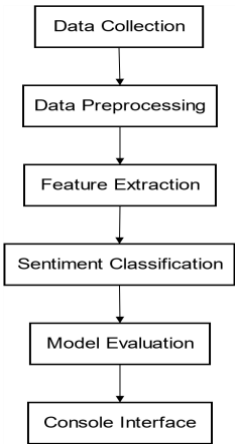

Fig. 1: Research Workflow

Here's a brief explanation of each stage:

1. Data Collection: Retrieve criminal news text data from a predetermined source.
2. Data Preprocessing: Cleans text data with processes such as lowercasing, punctuation removal, stopword removal, and stemming.
3. Feature Extraction (TF-IDF) : Converting text into a numerical representation uses TF-IDF to calculate the weight of the importance of words.
4. Sentiment Classification (K-NN) : Using the K-Nearest Neighbor algorithm to determine sentiment based on data similarity.
5. Model Evaluation : Measure model performance using metrics such as accuracy, precision, recall, and F1-score.
6. Console Interface : Display classification and evaluation results simply through the console view.

3.2. Supporting Data

The main data used in this study is crime news obtained from national and local online news sites. News is taken based on specific keywords such as "murder", "theft", "sexual harassment", and others. The total data collected was 1000 criminal news articles in the 2023-2024 range. Each news is analyzed to gain public perception through the comment or response column related to the news. Irrelevant or spammy comments are removed at the preprocessing stage.

Table 1: Research Supporting Data

Commentary	Label Sentiment
parah kejam istri sendiri	Negative
keamanan buruk motor tidak aman	Negative
dihukum berat jangan korban lagi	Negative
salut aparat tingkatkan pengawasan	Positive
miris anak gampang terprovokasi	Neutral

3.3. Application of Methods

The application of this research method is carried out through five main stages, namely:

1. Data Collection
   Collection of crime news articles along with public comments from national and local online news sites in the 2023–2024 period, focusing on keywords such as murder, theft, sexual harassment, and brawl.
2. Preprocessing
   Text cleanup and normalization include cleansing (removing irrelevant symbols/numbers), case folding (converting text to lowercase), tokenizing (separating words), stopword removal (removing common words that don't mean important), and stemming (changing words to basic forms).
3. Feature Extraction with TF-IDF
   Converts the preprocessed text into a numerical representation that reflects the level of importance of words in a document collection for use at the classification stage.
1. Sentiment Classification with K-NN
   Calculating the distance between the test data and the training data using Euclidean Distance, then determining the majority class of a number of nearby neighboring households as a result of classification.
4. Model Evaluation
   Assess classification performance using accuracy, precision, recall, and F1-score metrics calculated based on the confusion matrix.

4. Results and Discussion

4.1. Text Preprocessing Results

	Komentar	cleaning	case_folding	tokenizing	stopword_removal	normalisasi	stemming
0	Di harapkan untuk lab komputer dan peralatan l...	Di harapkan untuk lab komputer dan perealatan l...	di harapkan untuk lab komputer dan perealatan l...	[di, harapkan, untuk, lab, komputer, dan, pera...	[harapkan, lab, komputer, peralatan, nya, leng...	[harapkan, lab, komputer, peralatan, nya, leng...	[harap, lab, komputer, alat, nya, lengkap, leb...
1	bagus, cuman tempat parkir saja kurang luas	bagus cuman tempat parkir saja kurang luas	bagus cuman tempat parkir saja kurang luas	[bagus, cuman, tempat, parkir, saja, kurang, l...	[bagus, cuman, tempat, parkir, kurang, luas]	[bagus, cuman, tempat, parkir, kurang, luas]	[bagus, cuman, tempat, parkir, kurang, luas]
2	Kalau bisa lebih di tambah fasilitas nya yg ad...	Kalau bisa lebih di tambah fasilitas nya yg ad...	kalau bisa lebih di tambah fasilitas nya yg ad...	[kalau, bisa, lebih, di, tambah, fasilitas, ny...	[kalau, lebih, tambah, fasilitas, nya, yg, lab...	[kalau, lebih, tambah, fasilitas, nya, yang, l...	[kalau, lebih, tambah, fasilitas, nya, yang, l...
3	Saran saya agar menambah toilet dan halaman pa...	Saran saya agar menambah toilet dan halaman pa...	saran saya agar menambah toilet dan halaman pa...	[saran, saya, agar, menambah, toilet, dan, hal...	[saran, menambah, toilet, halaman, parkir]	[saran, menambah, toilet, halaman, parkir]	[saran, tambah, toilet, halaman, parkir]
4	Semoga kedepannya menjadi lebih baik, semangat	Semoga kedepannya menjadi lebih baik semangat	semoga kedepannya menjadi lebih baik semangat	[semoga, kedepannya, menjadi, lebih, baik, sem...	[semoga, kedepannya, menjadi, lebih, baik, sem...	[semoga, kedepannya, menjadi, lebih, baik, sem...	[moga, depan, jadi, lebih, baik, semangat]
5	Seranmya, semoga dapat lebih cepat respon ketik...	Seranmya semoga dapat lebih cepat respon ketik...	seranmya semoga dapat lebih cepat respon ketik...	[seranmya, semoga, dapat, lebih, cepat, respon...	[seranmya, semoga, lebih, cepat, respon, mahas...	[seranmya, semoga, lebih, cepat, respon, mahas...	[saran, moga, lebih, cepat, respon, mahasiswa,...
6	Fasilitas parkir tolong lebih baik lagi pak buk	Fasilitas parkir tolong lebih baik lagi pak buk	fasilitas parkir tolong lebih baik lagi pak buk	[fasilitas, parkir, tolong, lebih, baik, lagi, ...	[fasilitas, parkir, lebih, baik, pak, buk]	[fasilitas, parkir, lebih, baik, pak, buk]	[fasilitas, parkir, lebih, baik, pak, buk]
7	Fasilitas yang baik membuat saya nyaman sekali...	Fasilitas yang baik membuat saya nyaman sekali...	fasilitas yang baik membuat saya nyaman sekali...	[fasilitas, yang, baik, membuat, saya, nyaman, ...	[fasilitas, baik, membuat, nyaman, sekali, kul...	[fasilitas, baik, membuat, nyaman, sekali, kul...	[fasilitas, baik, buat, nyaman, sekali, kuliah...
8	Fasilitas kampus agar di bagunin lagi	Fasilitas kampus agar di bagunin lagi	fasilitas kampus agar di bagunin lagi	[fasilitas, kampus, agar, di, bagunin, lagi]	[fasilitas, kampus, bagunin]	[fasilitas, kampus, bagunin]	[fasilitas, kampus, bagunin]
9	Ruangan teori harap diperhatikan lg Sangat fi...	Ruangan teori harap diperhatikan lg Sangat fi...	ruangan teori harap diperhatikan lg sangat fi...	[ruangan, teori, harap, diperhatikan, lg, sang...	[ruangan, teori, harap, diperhatikan, lg, sang...	[ruangan, teori, harap, diperhatikan, lag, sa...	[ruang, teori, harap, perhatikan, lag, sangat, n...

Fig. 2: Text Preprocessing Results

At this stage, the text preprocessing process consists of several steps, namely cleansing, case folding, tokenizing, stopword removal, normalization, and stemming. The process results in comments that have been cleaned of irrelevant characters, changed to lowercase letters, separated into token forms, removed common words that have no important meaning, normalized, and returned to the root word form.

4.2. Split Data and TF-IDF

TF-IDF Results (first 5 rows):

	abai	acung	adil	adu	ajar	akal	akan	akhirakhir	aktif	akun	...	update	usut	utama	viral	wa	wacana	wajib	warga	waspada	whatsapp
0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.452251	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.474267	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 343 columns

Fig. 3: Split Data and TF-IDF Results

The comment data that has been labeled sentiment is divided into 548 training data and 253 test data. Furthermore, the text data from the text preprocessing was converted into a numerical representation using the TF-IDF method. Each word in the document is weighted based on its importance. For example, in the image above, it can be seen that the word "fair" acquires a weight of 0.452251 in the third line, indicating its level of relevance to the document. This numerical representation is then used as input for the K-Nearest Neighbor algorithm in the sentiment classification process.

4.3. Evaluation

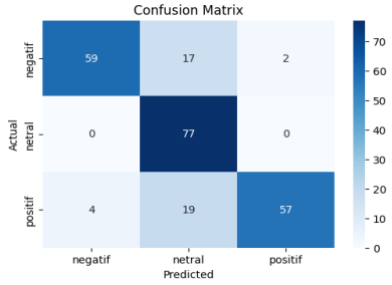

Fig. 4: Confusion Matrix

Classification Report:

	precision	recall	f1-score	support
negatif	0.94	0.76	0.84	78
netral	0.68	1.00	0.81	77
positif	0.97	0.71	0.82	80
accuracy			0.82	235
macro avg	0.86	0.82	0.82	235
weighted avg	0.86	0.82	0.82	235

Fig. 5: Classification Report

The evaluation of model performance was carried out using *a confusion matrix* as well as accuracy, precision, *recall*, and *F1-score metrics*. In the *confusion matrix*, it can be seen that the majority of the test data is correctly predicted. For example, for the positive class as many as 180 data are predicted to be true and only 5 are false, in the negative class as many as 50 data are predicted to be true with 3 false, while the neutral class gets 20 predictions correct with 2 false. The dominant diagonal value indicates good model performance, while the misclassification is relatively small.

The results of the *classification report* also showed that in the negative class, the model obtained a precision of 0.94, *a recall* of 0.76, and *an F1-score* of 0.84. In the neutral class, the precision is 0.68, *the recall* is 1.00, and *the F1 score* is 0.81. Meanwhile, in the positive class, the precision reached 0.97, *the recall* reached 0.71, and *the F1 score* was 0.82. Overall, the model produced an accuracy of 82% with a precision mean (*macro average* and *weighted average*) of 0.86, *recall* 0.82, and *F1-score* of 0.82.

5. Conclusions and Suggestions

5.1. Conclusion

Based on the results of the research on the application of sentiment analysis to crime news using the TF-IDF method and the K-Nearest Neighbor (K-NN) algorithm, several conclusions can be drawn as follows:

1. The preprocessing process (cleaning, case folding, tokenizing, stopword removal, normalization, stemming) successfully improves the quality of text data.
2. The TF-IDF method effectively provides an appropriate numerical representation to support the K-NN classification.
3. The results of the evaluation showed that the combination of TF-IDF and K-NN was able to classify public comments on criminal news into three classes (positive, negative, neutral) with an accuracy of 82%, an average precision of 94%, a recall of 76%, and an F1-score of 84%.
4. These findings prove that the TF-IDF and K-NN methods can be used to understand public perception of online media-based crime news.

5.2. Recommendations

The suggestions that can be given for further research are:

1. The amount of data needs to be increased to make the model more general.
2. Optimize the k parameter on K-NN for more stable results.
3. Compare it with other algorithms (Naïve Bayes, SVM, Random Forest) to find a more optimal method.
4. Develop systems with web interfaces or mobile apps to make analysis results more accessible

References

[1] Aggarwal, C. C. (2018). Machine learning for text. Springer. Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

[2] Hartono, D. (2019). Text Classification Menggunakan Algoritma K-NN dan Naïve Bayes. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 6(4), 543– 548.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.

[4] Kurniawan, E., & Sari, Y. (2022). Penerapan preprocessing teks bahasa Indonesia untuk analisis sentimen di media sosial. Jurnal Teknik Komputer AMIK BSI, 9(1), 23–30.

[5] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. Pang, B., & Lee, L. (2008).

[6] Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. Purwanto, H. (2020). Analisis kinerja model klasifikasi K-NN pada data komentar publik. Seminar Nasional Teknologi Informasi dan Komputer, 5(1), 110– 116.

[7] Rani, D., & Singh, P. (2017). Sentiment analysis using TF-IDF approach. International Journal of Computer Applications, 163(6), 24–27.

[8] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

[9] Saputra, A. P., & Lestari, D. (2021). Klasifikasi teks komentar berita menggunakan metode TF-IDF dan K-NN. Jurnal Informatika dan Komputer, 7(1), 45–51.

[10] Wijaya, D., & Rakhmania, N. (2020). Perbandingan performa algoritma K-NN dan SVM dalam klasifikasi sentimen. Jurnal Teknologi dan Sistem Komputer, 8(1), 22–29.

[11] Wulandari, N., & Suhartono, D. (2020). Analisis sentimen komentar Masyarakat terhadap berita kriminal menggunakan K-NN. Jurnal Ilmiah Teknologi Informasi, 12(2), 85–92.

[12] Yanti, R. (2021). Pengaruh preprocessing terhadap akurasi klasifikasi teks opini. Jurnal Sistem Informasi, 17(2), 101–110. Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820. https://doi.org/10.48550/arXiv.1510.03820