Journal of Science and Engineering (JSE), Vol. 1, No. 1 (page: 17-22)
e-ISSN: 2723-3944 September 2023
DOI: https://doi.org/10.30650/jse.v1i1.3773

Multilayer Perceptron and TF-IDF in the Classification of Hate Speech on Twitter in Indonesian

Akmal Syahrandi[1*], Asslia Johar Latipah[2], Naufal Azmi Verdikha[3]
[1,2,3] Universitas Muhammadiyah Kalimantan Timur, Samarinda, Indonesia
Email* : 1911102441086@umkt.ac.id

Abstract – Twitter nowadays is one of the popular social media which currently has over 300millions accounts, twitter is the rich source to learn about people's opion and sentimental analysis. However, this also brings new problems where the practice of hate speech. This research classifies of hate speech on social media. Evaluation using dataset from previous research Ibrohim&Budi (2019), then using classification method Multilayer Perceptron which combined with feature extraction to be able to detect negations and weighting uses Term Frequency – Inverse Document Frequency (TF-IDF). Results show that the F1 score gives an accuracy rate of up to 74.51%. This research has a reasonably good effectiveness from combining the TF-IDF and Multilayer Perceptron methods, considering the results obtained from the F1 Score evaluation value.

Keywords: Hate Speech, Multilayer Perceptron, TF-IDF

Submitted : 27 Juli 2023 - Revised : 30 Agustus 2023 - Accepted : 18 September 2023

1. Introduction

The usage of social media, tends to grow each year, has given rise to a new phenomena. The Indonesian Internet Service Providers Association (APJII) conducted a poll of 2022 internet users and found that 89.15% of them frequently browse social media. In comparison to online socializing, talking, shopping, and other activities, this percentage is large. Social media technology gives people the ability to share their opinions, including hate speech, that subsequently spreads widely and becomes viral if the issues covered are "interesting." This could lead to conflicts between groups on social media. This also creates new issues because hate speech is becoming more common through this media[1]. It demonstrates how the public social media space, which was intended to serve as a forum for the exchange of information, concepts, and knowledge, has changed into a setting for the dissemination of hate speech texts, disrespectful sentences, such as insults and curse words, that make it difficult for netizens to communicate effectively and creates hostility[2].

According to the National Police Criminal Investigation Agency of Indonesia, 143 hate speech-related cybercrimes were committed in Indonesia in 2015[3]. One of them is social media Twitter. Twitter nowadays is one of the popular social media currently has over 300millions accounts, twitter is the rich source to learn about people's opion and sentimental analysis[4]. According to data from We Are Social, Twitter occupies the fifth position of social media often used by the people of Indonesia. In the survey, it was seen that as many as 63.6%, or equivalent to 108 million Indonesians aged 16 to 64, are users who spend their time using Twitter[5].

One of the methods that researchers have attempted to automatically reduce hate speech is by using machine learning algorithms. Differences in opinion among people who decide whether to classify a piece of texts as hate speech or not present some challenges in identifying hate speech. It shows the potential for misclassification in machine learning algorithms that will later be trained based on human labeling[6].

Previous research by Ibrohim & Budi (2019) has successfully created datasets to identify abusive language and hate speech in Indonesian on the Twitter platform. The dataset is based on crawling results on the Twitter platform. Experiments were also conducted using unigram words, Random Forest Decision Tree (RFDT), and Label Power-set (LP) methods as the best combination of features, classifiers, and data transformation methods for all scenarios performed. Based on the experiments, an accuracy rate of 77.36% was produced to perform multi-label classification to identify abusive language and hate speech without identifying the target, category, and level of hate speech. On the other hand, 66.12% at the time of conducting multi-label classification to identify abusive language and hate speech, including identifying targets, categories, and levels of hate speech.

Using different methods, research by Polignano and

Basile (2018) has compared the classification methods of Logistic Regression, Support Vector Classification, K-nearest neighbors, Decision Tree, Random Forest, and Multilayer Perceptron classifier with TF-IDF featured extraction for Italian hate speech classification[7]. The final result of the study was that Multilayer Perceptron achieved an F1 Score of 79.1%. These results prove that Multilayer Perceptron has a better F1 Score value compared to the Logistic Regression algorithm with a value of 78%, Support Vector Classification with a value of 78.9%, K-nearest neighbors with a value of 70.5%, Decision Tree with a value of 68.0%, and Random Forest with a value of 78.7%.

With the background outlined, this study will identify hate speech on social media Twitter. This research adopts the method used by Polignano and Basile, namely by using Multilayer Perceptron combined with TF-IDF as its extraction feature. This study uses a dataset made by Ibrohim & Budi (2019), the F1 Score evaluation value will be measured. Hopefully, this study's results can significantly contribute to addressing the problem of hate speech on social media and improve understanding of natural language processing and machine learning techniques in text analysis.

2. Related Works

The implementaion of Multilayer Perceptron was applied to Amalia Y. & Sibaroni Y. research (2020). Researchers analysis about sentimen Tweet data on the planned relocation of the Indonesian capital. with weighting using TF-IDF. The results of the built model are delta TF-IDF obtained the highest accuracy result of 70.6%. And TF-IDF at 68.5%[8]

The use of TF-IDF and Multilayer Perceptron method in Muzakki, Jondri & Umbara (2019). Researchers analysis about Student Questionnaire on Telkom University facilities. The Accuracy value uses the best Confusion Matrix calculation with a percentage of 91.23%[9].

Subsequent research uses the Multilayer Perceptron method for Tweet data sentiment of Indonesian netizens on the COVID-19 Vaccine. The classification model results with 68.8% accuracy, 0.82 precision, and 0.64 recall[10].

3. Research Method

Research methodology is a process that is required in research. In this research, the following (Figure 1) is an overview of the sequence of stages to be carried out:

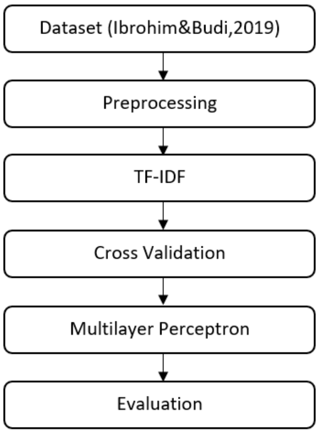
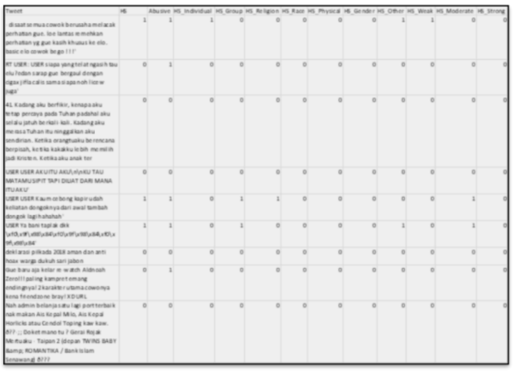


Fig. 1. Research Method

Further explanation regarding the research stage in Figure 1 is as follow:

1. Using Dataset from Ibrohim&Budi (2019).
2. Preprocessing to prepare text data before it is used in next processes.
3. TF-IDF as feature extraction which is implemented using Scikit-Learn.
4. Cross Validation to spilt dataset to data training and data test.
5. Multilayer Perceptron as a classification which is implemented using Scikit-learn.
6. Evaluation with F1 Score to evalution the model.

This research uses data from This study uses a dataset from previous research (Ibrohim & Budi, 2019) on the GitHub site. The reference dataset uses multi-label information to identify offensive language and hate speech. The data will be used in this study, the following is an example of the top 10 dataset shown in Fig. 2. Dataset with "Tweet" column visualized in *Wordcloud* form using *matplotlib* library shown in Fig. 3.



Tweet	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
- disaat semua cowok berusaha melacak perhatian gue. loe lantas remehkan perhatian yg gue kasih khusus ke elo. basic elo cowok bego ! ! !'	1	1	1	0	0	0	0	0	1	1	0	0
RT USER: USER siapa yang telat ngasih tau elu?edan sarap gue bergaul dengan cigax jifla calis sama siapa noh licew juga'	0	1	0	0	0	0	0	0	0	0	0	0
41. Kadang aku berfikir, kenapa aku tetap percaya pada Tuhan padahal aku selalu jatuh berkali-kali. Kadang aku merasa Tuhan itu ninggalkan aku sendirian. Ketika orangtuaku berencana berpisah, ketika kakakku lebih memilih jadi Kristen. Ketika aku anak ter	0	0	0	0	0	0	0	0	0	0	0	0
USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT TAPI DILIAT DARI MANA ITU AKU'	0	0	0	0	0	0	0	0	0	0	0	0
USER USER Kaum cebong kapir udah kelihatan dongoknya dari awal tambah dongok lagi hahahah'	1	1	0	1	1	0	0	0	0	0	1	0
USER Ya bani taplak dkk \xf0\x9f\x98\x84\xf0\x9f\x98\x84\xf0\x9f\x98\x84'	1	1	0	1	0	0	0	0	1	0	1	0
deklarasi pilkada 2018 aman dan anti hoax warga dukuh sari jabon	0	0	0	0	0	0	0	0	0	0	0	0
Gue baru aja kelar re-watch Aldnoah Zero!!! paling kampret emang endingnya! 2 karakter utama cowonya kena friendzone bray! XD URL	0	1	0	0	0	0	0	0	0	0	0	0
Nah admin belanja satu lagi port terbaik nak makan Ais Kepal Milo, Ais Kepal Horlicks atau Cendol Toping kaw kaw. 877 ;;; Doket mano tu ? Gerai Rojak Mertuaku - Taipan 2 (depan TWINS BABY &amp; ROMANTIKA / Bank Islam Senawang) 877?	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 2. Top 10 Dataset

Fig. 3. Wordcloud “Tweet” column

Fig.3 displays words often appearing in the "Tweet" column. Based on the frequency, it can be seen that some tweets from the dataset have words that have no meaning. Pre-processing steps are needed before implementing Machine Learning, it is necessary to remove noise or distractions, normalize data that is not normal, and also pre-processing is needed to reduce the size of data that is too large, because data that is too large can become a distraction because there is something unnecessary/relevant in doing Machine Learning. The pre-processing stages that have been carried out as follows:

- Insert Cloumn ID: By using ID columns, the process can have a consistent and unique way to identify, organize, and manage data in datasets, and perform operations and manipulation of data more efficiently and accurately.
- Input dictionary “kamus_alay” and stopword: At this stage, will input two dictionaries called "kamus alay", and “kamusalay_clean5”. And at this stage will input a stopword. Dictionary “kamus alay” uses the dictionary that has been provided by Ibrohim&Budi (2019), and dictionary “kamusalay_clean5” made to add lack of word form previous dictionary. This stage is intended as a reference for normalizing words, spelling, and removing irrelevant words that will be used in the next step.
- Lowercase: In the data used in this study, there are still uppercase and lowercase letters.
- Remove Attribute Tweet: removing certain attributes from tweets that are irrelevant or not needed for analysis
- Non-Alphanumeric: to remove punctuation or non-alpha numeric characters contained in research data.
- Spell Checker: checks for word errors or incorrect spelling. It provides alternatives to the correct word or spelling by using the previously inputted library.
- Stemming: With the stemming process, this stage will remove the initial or final affix to the word.
- Stopword Removal: remove words that have no meaning or value in the sentence.

The TF-IDF method is a method for calculating the weight of each word that is most commonly used in information retrieval. This method is also known to be efficient, easy and has accurate results[11].

The establishment of the TF-IDF extraction feature in determining document text can be exemplified in a simple manner as follows:

Table 1
TF-IDF Sample Text

No	Text
D1	Saya pergi
D2	Saya bekerja
D3	Saya pergi bekerja

From the text of the statement above it can be arranged into a TF-IDF. His intuition is that a word appearing in multiple documents is not a good differentiator and should be given less weight than one that occurs in multiple documents. Merging Term Frequency (TF) schemes with Inverse Document Frequency (IDF) has proven to be a powerful method for processing text data or other purposes[12].

Table 2
TF-IDF Calculation Result

TF-IDF	Saya	Pergi	bekerja
D1	1	1.2877	0
D2	1	0	1.2877
D3	1	1.2877	1.2877

This study uses Scikit Library (sklearn) for feature extraction using TF-IDF. In Sklearn, the TF-IDF extraction feature can be implemented using the fidfVectorizer module.

After extracting the text data, the dataset is divided randomly into training data and test data using Cross-Validation. One of cross-validation technique is K-fold cross validation, which breaks data into K pieces of the dataset of equal size, and using K-fold cross-validation to eliminate bias in data. Training and testing are carried out k times. In the first experiment, the S1 subset is treated as test data, and the other subset is treated as training data, in the second attempt, the S1, S3 subset... Sk becomes training data, S2 becomes testing data, and so on[13].

This research uses K-fold Cross Validation with 10-Fold Cross Validation using the library from scikit library. The first fold will be used as test data in the first test, and the remaining nine will be used as train data. An overview of the process can be seen below (fig. 4) for 10 folds:

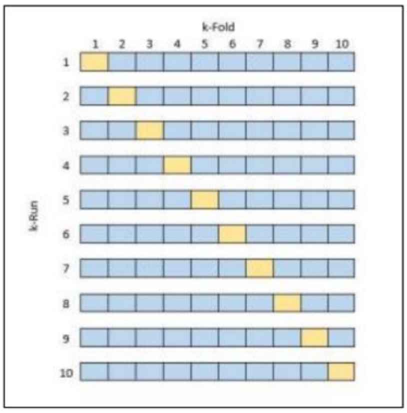


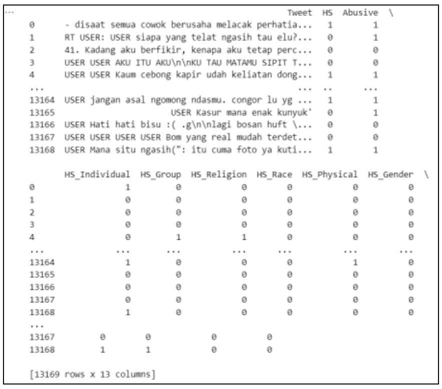
Fig. 4. 10-Fold Cross-Validation

The classification and evaluation used are Multilater Perceptron and F1 Score. Multilayer Perceptrons are a class of feedforward neural networks built by layered acyclic graphs. A Multilayer Perceptron consists of at least three layers and non-liner activation. The first layer is called the input layer, the second layer is called the hidden layer, and the third layer is called the output layer. All three layers are fully connected, meaning every node in the hidden layer is connected to every node in the other layers. Multilayer perceptrons are trained using backpropagation, where weights are updated by calculating gradient decreases for the error function[14].

There are several ways to measure classification methods' performance, including the Confusion Matrix, Precision, Recall, and F1-Score. F1-Score is an evaluation calculation in retrieval information that combines recall and precision. Precision is the ratio of the correct categorization of documents into categories to the total number of classification attempts. Recall is the level of the system's ability to recover relevant information[15].

4. Result and Discussion

The data used in this study is called *Comma Separated Values* (CSV). By using the read.csv function, the result data that has been imported into Python is shown in the following fig.5:



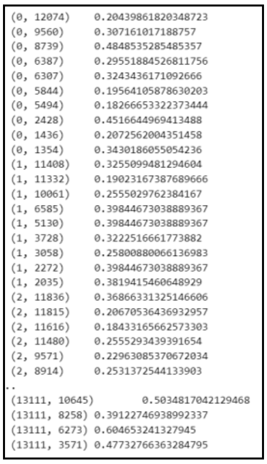
```
...                                                   Tweet  HS  Abusive  \
0      - disaat semua cowok berusaha melacak perhatia...   1        1
1      RT USER: USER siapa yang telat ngasih tau elu?...   0        1
2      41. Kadang aku berfikir, kenapa aku tetap perc...   0        0
3      USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...   0        0
4      USER USER Kaum cebong kapir udah keliatan dong...   1        1
...                                                  ...  ..      ...
13164  USER jangan asal ngomong ndasmu. congor lu yg ...   1        1
13165                       USER Kasur mana enak kunyuk'   0        1
13166  USER Hati hati bisu :( .g\n\nlagi bosan huft \...   0        0
13167  USER USER USER USER Bom yang real mudah terdet...   0        0
13168  USER Mana situ ngasih(": itu cuma foto ya kuti...   1        1

       HS_Individual  HS_Group  HS_Religion  HS_Race  HS_Physical  HS_Gender  \
0                  1         0            0        0            0          0
1                  0         0            0        0            0          0
2                  0         0            0        0            0          0
3                  0         0            0        0            0          0
4                  0         1            1        0            0          0
...              ...       ...          ...      ...          ...        ...
13164              1         0            0        0            1          0
13165              0         0            0        0            0          0
13166              0         0            0        0            0          0
13167              0         0            0        0            0          0
13168              1         0            0        0            0          0
...
13167         0        0           0          0
13168         1        1           0          0

[13169 rows x 13 columns]
```

Fig. 5. Preparing Dataset

After successfully importing data from previous studies, the TF-IDF process was carried out. The following is the output of the BoW data extract:



```
(0, 12074)    0.20439861820348723
(0, 9560)     0.307161017188757
(0, 8739)     0.4848535285485357
(0, 6387)     0.29551884526811756
(0, 6307)     0.3243436171092666
(0, 5844)     0.19564105878630203
(0, 5494)     0.18266653322373444
(0, 2428)     0.4516644969413488
(0, 1436)     0.2072562004351458
(0, 1354)     0.3430186055054236
(1, 11408)    0.3255099481294604
(1, 11332)    0.19023167387689666
(1, 10061)    0.2555029762384167
(1, 6585)     0.39844673038889367
(1, 5130)     0.39844673038889367
(1, 3728)     0.3222516661773882
(1, 3058)     0.25800880066136983
(1, 2272)     0.39844673038889367
(1, 2035)     0.3819415460648929
(2, 11836)    0.36866331325146606
(2, 11815)    0.20670536436932957
(2, 11616)    0.18433165662573303
(2, 11480)    0.2555293439391654
(2, 9571)     0.22963085370672034
(2, 8914)     0.2531372544133903
..
(13111, 10645)        0.5034817042129468
(13111, 8258) 0.39122746938992337
(13111, 6273) 0.604653241327945
(13111, 3571) 0.47732766363284795
```

Fig. 6. Result of TF-IDF

The output results has been calculated with TF-IDF, namely (0, 12074), indicating that in corpus 0 there is the 12074th index with a value of 0.204 and so on until the end, namely (13111, 3571) which indicates that in corpus 13111 there is the 3571st index with a value of 0.477.

After extracting feature of the text data, the next step is to split the dataset into training data or training and test data with Cross-Validation. The results/output of the cross-validation with example on Fold-1 can be seen in the following table.

Table 3.
Fold-1 Cross Validation

Indeks of Test Data	Indeks of Train Data
0, 1, 2, ..., 1309, 1310, 1311.	1312, 1313, 1314, ..., 13109, 13110, and 13111

Table 3 above shows that in Fold 1, are 1312 data used to test the model (testing), with indices 0, 1, 2, ..., 1309, 1310, and 1311. There are also 11800 data used to train the model (training), with indices 1312, 1313, 1314, ..., 13109, 13110, and 13111. The cross-validation results on Fold 1 follow the Fig. 4 described in the previous chapter, which shows that this applies to the next fold up to Fold 10.

The next stage is forming the Multilayer Perceptron model using the import MLPClassifier from the scikit-learn library. After the model is formed, cross-validation of the features is carried out using the cross_validation function with a k-fold of 10. Then, the classification results are evaluated by calculating the F1 Score. The results of this phase can be seen in the following Fig. 7:

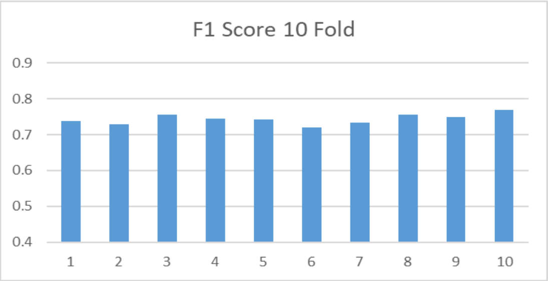


Fig. 7. F1 Score 10-Fold

It can be seen that the highest F1 score is in the 10th fold. That shows that it has good effectiveness at the cross-validation stage with the train data and test data in the fold. So that the value obtained at the 10th fold produces a high F1 score among the other folds. The following is the result of the overall fold on Confusion Matrix:

Table 4.
Overall Score Confusion Matrix

Confusion Matrix	Overall Score
True Positive	4055
False Positive	1291
True Negative	6270
False Negative	1496

Table() shows the overall fold on the confusion matrix, with overall scores are 4055 as identified for true positive, 1291 for false positive, 6270 for true negative, and 1496 for false negative. The values obtained are used to calculate the F1 score. The evaluation results result from using TF-IDF and Multilayer Perceptron in classifying hate speech. The following are the results of the F1 Score obtained:

```
F1 Score: 0.7442
```

Figure 8. F1 Score result

Figure() show the result of an average F1 score with a result of 0.7442. These results indicate the level of accuracy of the classification model used in identifying hate speech. With the results that have been obtained, the model that has been done has a relatively good level of precision and recall in classifying. These results indicate that the model has good effectiveness in detecting hate speech.

5. Conclusion

Based on the research results that have been obtained and discussed previously, it can be concluded that the results of combined the TF-IDF method with the Multilater Perceptron with F1 Scores get an average of 0.7742. The score that has been obtained shows that the class prediction results are relatively accurate, along with the results of the confusion matrix that have been obtained.

This research has a reasonably good effectiveness from combining the TF-IDF and Multilayer Perceptron methods, considering the results obtained from the F1 Score evaluation value, where if it is close to 1, then the model.

References

[1] C. Juditha, “Hatespeech In Online Media: Jakarta On Election 2017-Hatespeech di Media Online: Kasus Pilkada DKI Jakarta 2017,” *J. Penelit. Komun. Dan Opini Publik*, pp. 137–151, 2017.

[2] U. Ulinnuha and M. Ulum, “Efektivitas Pembelajaran Bahasa Indonesia bagi Mahasiswa dalam Menghindari Ujaran Kebencian di Media Sosial,” *IKRA-ITH Hum. J. Sos. dan Hum.*, vol. 6, no. 3, pp. 12–23, 2022, doi: 10.37817/ikraith-humaniora.v6i3.2119.

[3] M. A. Fauzi and A. Yuniarti, “Ensemble method for indonesian twitter hate speech detection,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 11, no. 1, pp. 294–299, 2018, doi: 10.11591/ijeecs.v11.i1.pp294-299.

[4] Y. Wang, J. Guo, C. Yuan, and B. Li, “Sentiment Analysis of Twitter Data,” *Appl. Sci.*, vol. 12, no. 22, 2022, doi: 10.3390/app122211775.

[5] K. Simon, “DIGITAL 2021: THE LATEST INSIGHTS INTO THE ‘STATE OF DIGITAL,’” *wearesocial.com*, 2021. https://wearesocial.com/uk/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/ (accessed May 19, 2023).

[6] G. B. Herwanto, A. Maulida Ningtyas, K. E. Nugraha, and I. Nyoman Prayana Trisna, “Hate Speech and Abusive Language Classification using fastText,” *2019 2nd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2019*, pp. 69–72, 2019, doi: 10.1109/ISRITI48646.2019.9034560.

[7] M. Polignano and P. Basile, “Hansel: Italian hate speech detection through ensemble learning and deep neural networks,” *CEUR Workshop Proc.*, vol. 2263, 2018, doi: 10.4000/books.aaccademia.4766.

[8] C. Amalia and Y. Sibaroni, “Analisis Sentimen Data Tweet Menggunakan Model Jaringan Saraf Tiruan Dengan Pembobotan Delta Tf-idf,” *eProceedings …*, vol. 7, no. 2, pp. 7810–7820, 2020, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/12799

[9] M. F. Muzakki, R. F. Umbara, F. Informatika, and U. Telkom, “Analisis Sentimen Mahasiswa Terhadap Fasilitas Universitas Telkom Menggunakan Metode Jaringan Saraf Tiruan Dan Tf-Idf,” *e-Prodeceeding Eng.*, vol. 6, no. 2, pp. 8608–8616, 2019.

[10] C. Lestari, “Sentiment Analysis Pandangan Netizen Indonesia Terhadap Vaksin COVID-19,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 4, pp. 2795–2803, 2022, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/2518%0Ahttps://jurnal.mdp.ac.id/index.php/jatisi/article/download/2518/1006

[11] S. Robertson, “Understanding inverse document frequency: On theoretical arguments for IDF,” *J. Doc.*, vol. 60, no. 5, pp. 503–520, 2004, doi: 10.1108/00220410410560582.

[12] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” *citeseerx.ist.psu.edu*, 2003.

[13] F. Tempola, M. Muhammad, and A. Khairan, “Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.

[14] A. Nyberg, “Classifying movie genres by analyzing text reviews,” pp. 1–12, 2018, [Online]. Available: http://arxiv.org/abs/1802.05322

[15] G. Forman, “10.1162/153244303322753670,” *CrossRef List. Deleted DOIs*, vol. 1, pp. 1289–1305, 2000, doi: 10.1162/153244303322753670.