

Applying TF-IDF and K-NN for Clickbait Detection in Indonesian Online News Headlines

Muhammad Athallah Afif¹, Munirul Ula², Lidya Rosnita^{3*}, Rizal⁴

^{1, 2, 3, 4} Universitas Malikussaleh, Indonesia

*Corresponding Author Email: lidyarosnita@unimal.ac.id

ABSTRACT

Received: 23 March 2024

Revised: 31 March 2024

Accepted: 31 March 2024

Available online: 1 April 2024

Keywords:

TF-IDF; k-Nearest Neighbor; Clickbait; Online News Headlines; Indonesian

This research explores the application of TF-IDF (Term Frequency-Inverse Document Frequency) and K-Nearest Neighbor (K-NN) in constructing a clickbait detection system for Indonesian online news headlines. The TF-IDF method is employed to ascertain the significance of words in news headlines, utilizing a tokenization process to generate numeric representations. The TF-IDF matrix serves as features in the K-NN classification model, with $k=1$ determining the most similar class. Model evaluation yields outstanding results, achieving accuracy, precision, recall, and F1-Score all reaching 1.0. The confusion matrix unveils no misclassifications, affirming the model's adeptness in correctly classifying all samples.

1. INTRODUCTION

In the era of Industry 4.0, the progress of knowledge and innovation is advancing rapidly, particularly in the realm of web technology. Almost every region in the world is leveraging these innovations to search for and disseminate data. Similarly, a significant portion of the Indonesian population uses the internet to gather and distribute information. According to the 2018 APJII survey, Java and Sumatra still have the highest percentage of internet users. Based on survey findings, a web news portal will be created exclusively to provide agricultural-related news through widespread internet use. Here, agriculture refers to human activities utilizing biological resources to produce food, raw materials for industries, or sources of energy [1]. The online news portal has become one of the mass media with significant power in disseminating information. A common complaint among users is the difficulty in understanding the portal's interface and functionality that has not yet been fully utilized [2]. An online news portal must meet several aspects to maintain the quality of user experience and the provided information, one of which is usability. Usability determines the level of usefulness of the news portal, assessed based on effectiveness, efficiency, and satisfaction through usability evaluation [2].

Clickbait is content titles created to grab attention and entice visitors to click on a specific web page link. Clickbait is also referred to as trap links in content titles designed to capture the reader's attention. However, the content itself is often ordinary and sometimes unrelated to the title. Clickbait practices are commonly found in online news and social media content [3]. According to [4] in the research titled "Clickbait News Classification Using K-Nearest Neighbor," one way to increase reader and visitor traffic is through clickbait. This deceptive content practice impacts news site

providers due to user interest and the difficulty users face in distinguishing misleading content from non-misleading news content. Deceptive content actions heavily depend on news site providers who often use sensational titles to attract users. To distinguish between deceptive and non-deceptive news content, K-Nearest Neighbors (K-NN) is employed, where KNN processing time is faster compared to other methods. From the examination consequences leading to the sequence of misleading news content, the best results were obtained at $k = 11$ involving situation 1 with 800 preparation information and 200 test information, delivering accuracy of 71%, precision of 72%, and recall of 71%. This indicates that the sequence of misleading news content can be characterized using K-Nearest Neighbor. The most commonly used technique for assigning weights to a term is TF-IDF. The rationale behind using this weighting is to assign value to a term, where the value of the term serves as a contribution to the clustering process [5]. K-Nearest Neighbor is a method that enables the classification of data based on its nearest distances. Additionally, K-NN falls under the category of supervised learning algorithms, where the learning process relies on predictor variable values. In the K-NN algorithm, all data must have labels, so when new data is provided, a comparison is made with existing data, and the most similar data is then selected based on its label [6].

Aided by the fact that the Indonesian society, in general, is more susceptible to the impact of misleading content, reflected by articles or content with misleading titles having higher average access compared to articles using non-misleading news titles. Strengthening the author's purpose behind this creation is the premise examined in "Application of TF-IDF and K-Nearest Neighbor in Building Clickbait Detection System in Indonesian Online News Headlines."

2. RESEARCH METHODS

The following is a flowchart that illustrates the steps in this research:

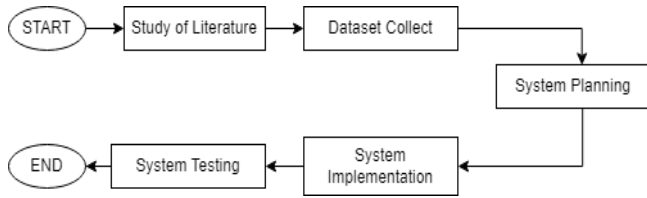


Figure 1. Research Flowchart

a. Study of Literature

In the effort to search for and collect data and information related to this research, the author conducted a literature review involving topics such as online news portals, clickbait, Term Frequency-Inverse Document Frequency (TF-IDF), and k-Nearest Neighbor. The literature review process was carried out by referring to various sources, including research journals, websites, and e-books. All references used in this research will be listed in the bibliography.

b. Dataset Collect

Data collection in this research is conducted systematically, focusing on online news headlines in the Indonesian language. The data used is sourced from reliable outlets that provide a set of news headlines labeled as clickbait and non-clickbait. This dataset includes various language variations and writing styles commonly encountered in online news. Each news headline is collected along with the corresponding label to train and test the developing clickbait detection system. The selection of a diverse dataset is crucial to ensure accurate and valid research results. The data collection process also involves preprocessing steps to ensure consistency and cleanliness of the data before integration into the clickbait detection system. These steps are crucial to ensure reliable and relevant analysis results regarding online news headlines in the Indonesian language.

c. System Planning and Implementation

In creating and utilizing this system, the author focuses on a responsive user interface design, using HTML and CSS to create a clean and user-friendly input form. Users can input the news headlines they want to analyze. Next, the author reads the news dataset and performs tokenization, transforming words into tokens, to prepare the data for the next stage.

For TF-IDF calculation, the author uses the TF-IDF Vectorizer library from sklearn to provide a numerical representation of each news headline. The system also involves the implementation of the K-NN model (K-Neighbors Classifier) for clickbait detection. This model is trained using the TF-IDF matrix and labels for clickbait or non-clickbait. The detection results and similarity scores between new headlines and clickbait/non-clickbait categories are then displayed on the user interface.

The importance of system testing is a concern for the author during implementation. The author ensures that the system is tested with various test cases covering both clickbait and non-clickbait news headlines. Similarity scores between new headlines and clickbait/non-clickbait categories

are displayed to provide additional information to the user. Afterward, the application is integrated with Flask and launched on port 8000, providing users with ease in detecting clickbait in Indonesian news headlines. With these steps, it is expected that the system can provide accurate and reliable detection results. Subsequently, the research will focus on evaluating the results and discussing the findings generated by the system.

d. System Testing

Flowchart of the testing process for the clickbait detection system using the TF-IDF method and the k-Nearest Neighbor Algorithm:

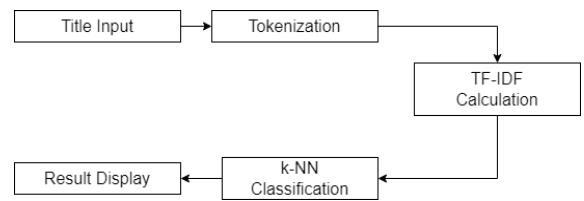


Figure 2. System Testing Flowchart

3. RESULT AND DISCUSSION

a. Initial Data Structure

Initially, the data structure in this research consists of a CSV-formatted dataset stored in the "dataset.csv" file. This dataset includes online news headlines in the Indonesian language, along with labels indicating whether the headlines are classified as clickbait or non-clickbait. Here is an example of the data structure in the "dataset.csv" file:

Table 1. Example Dataset Structure in Tabular Form

No	Title	Label	Label_score
1	Kehadiran Menantu Jokowi di Pilwalkot Medan, Jadi Sorotan Radar	non-clickbait	0
2	Malaysia Membahas Isu Kabut Asap dan Tuduhan Invasi Babi Oleh RI	non-clickbait	0
3	Driver Ojol di Bekasi Viral Antar Pesanan Pakai Sepeda	clickbait	1
4	Bantuan Rp 7,3 M dari Kemensos untuk Korban Kerusakan di Papua	non-clickbait	0
5	Pria Ditangkap terkait Mayat Bayi di Tangerang	non-clickbait	0
6	Raibnya Uang Rp 1,6 M di Parkiran Kantor Gubernur Sumut	non-clickbait	0
7	MPR: Amandemen UUD 1945 Tak Akan Membuat Perubahan Signifikan	non-clickbait	0
8	Munculnya Istana di Tengah Upaya Rekonsiliasi di Rusuh Papua	non-clickbait	0
9	Festival Muharam di Banyuwangi untuk Memperingati Tahun Baru Islam	non-clickbait	0
10	Wanita Asal Kendari Hampir Jadi Korban Pencobaan Perkosa di Makassar	non-clickbait	0

This dataset serves as the foundation for training the model, algorithm, and conducting analyses on the clickbait detection system. The process of reading and understanding

the dataset structure is carried out through the 'read_dataset' function in the 'main.py' program.

b. System and Model Performance

At this stage, the performance evaluation of the clickbait detection system is conducted based on the generated K-Nearest Neighbor (K-NN) model. This evaluation provides an overview of how well the system can recognize clickbait and non-clickbait based on the TF-IDF representation in Indonesian news headlines. The steps in evaluating the system's performance involve several commonly used evaluation metrics, as follows:

1. Accuracy

Model accuracy measures how well the model can classify data correctly.

In this case, the model produces a value of: 1.0.

The model accuracy result indicates that the model has the ability to classify news headlines as clickbait or non-clickbait.

2. Precision

Precision measures how well positive results are generated by the model.

In this case, the model produces a value of: 1.0.

The precision value indicates that the model tends not to produce many false positives when classifying news.

3. Recall

Also known as sensitivity or true positive rate, recall measures how well the model can identify all instances that should be positive.

In this case, the model produces a value of: 1.0.

The recall value indicates that the model is capable of identifying most of the actual clickbait headlines.

4. F1-score

F1-Score is a comparison between precision and recall.

In this case, the model produces a value of: 1.0

This F1-Score indicates that the model has a good balance between precision and recall.

5. Confusion Matrix

The Confusion Matrix presents the model's prediction results on the data and provides further insights into the model's performance.

All the data described above can be seen in the following tables:

Table 2. Evaluation Matrix Results

	Accuracy	Precision	Recall	F1-Score
Nilai	1.0	1.0	1.0	1.0

Table 3. Confusion Matrix Results

	Predicted 0	Predicted 1
Actual 0	110	0
Actual 1	0	890

c. Implementation Program Results

The steps to use this program are as follows:

1. Prepare the environment

As the author did not host this program for public use, localhost is used to run the program. Before using this program, it is essential to install 'Flask' in the local environment.

2. Download and Prepare the Dataset

Certainly, in using this model, the author requires a dataset. The dataset used by the author is obtained from the Mendeley website. The dataset is named 'dataset.csv' with a .csv file extension, and it should be placed in a specific directory along with the main code program, which is 'main.py'. This dataset includes news headlines, clickbait and non-clickbait labels, and score labels indicating whether the headline is classified as clickbait or not.

3. Run Flask

Open the command prompt terminal, navigate to the 'ClickbaitDetector' directory, and run Flask using the command 'python main.py'. If Flask is already running in the command prompt console, open a browser and enter the local address of the program, which is localhost with port 8000, as follows: 'localhost:8000'.

4. Use Clickbait Detector

After entering 'localhost:8000', the user interface of the Clickbait Detector program will be displayed. The following image is accompanied by instructions on how to use the Clickbait Detector program:

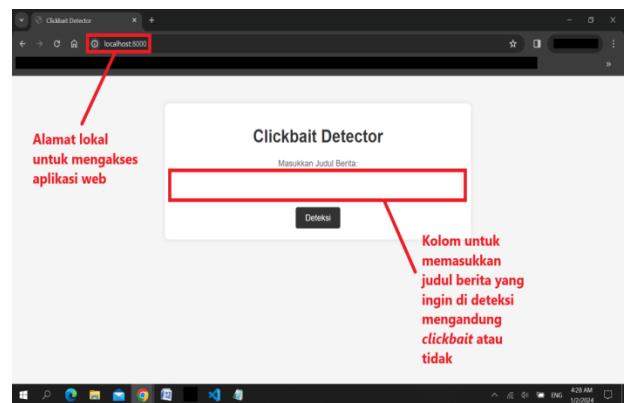


Figure 3. Input Page

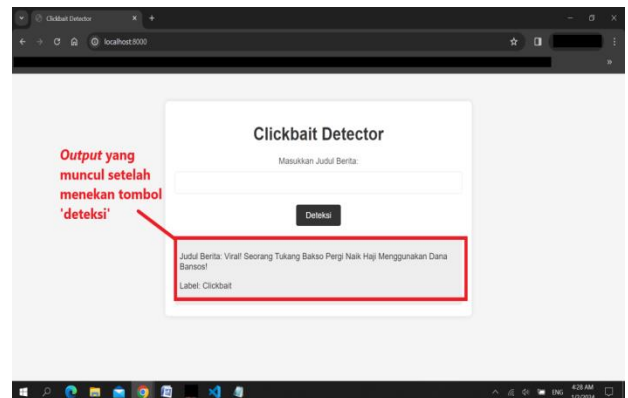
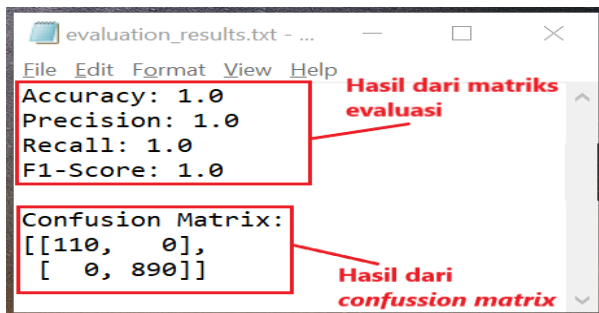


Figure 4. Output Page

5. Model Evaluation

The results of the model evaluation, including accuracy, precision, recall, F1-score, and confusion matrix, can be found in the file automatically generated after running the program. The file is located in the same directory as the main code program 'main.py' and is automatically named 'evaluation_results' with the extension '.txt'. The contents of the file can be seen in the following image:



```
evaluation_results.txt - ...
File Edit Format View Help
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1-Score: 1.0
Confusion Matrix:
[[110, 0],
 [ 0, 890]]
```

Figure 5. Evaluation Results Output from The Program

4. CONCLUSIONS

This research discusses the implementation of Term Frequency-Inverse Document Frequency (TF-IDF) and the K-Nearest Neighbor (K-NN) algorithm in building a clickbait detection system for Indonesian online news headlines. The TF-IDF method is employed to determine the significance of words in the headlines, followed by tokenization to generate numerical representations. The TF-IDF matrix is then utilized as features in the classification model, employing the K-NN algorithm to distinguish between clickbait and non-clickbait titles. In using the K-NN model, a k value of 1 is chosen to determine the most similar class. Model evaluation results demonstrate excellent performance, with accuracy, precision, recall, and f1-score reaching 1.0. The confusion matrix shows no misclassifications, indicating that the model successfully classifies all samples correctly.

REFERENCES

[1] Y. Devianto and S. Dwiasnati. (2021). "Rancang Bangun Web Portal Berita Sebagai Sumber Informasi

Berita Tentang Pertanian," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 534–546.

[2] P. T. Fajarini, N. Kadek, A. Wirdiani, and I. P. A. Dharmaadi. (2020). "Evaluasi Portal Berita Online Pada Aspek Usability Online Portal Evaluation On Usability Aspect Using Heuristic," vol. 7, no. 5, pp. 905–910.

[3] A. Daradinanti and V. Karunia Mulia Putri, "Clickbait: Pengertian, Jenis, dan Contohnya," *kompas.com*, 2022. [Online]. Available: <https://www.kompas.com/skola/read/2022/05/18/103000469/clickbait--pengertian-jenis-dan-contohnya?page=all>. Accessed on Dec. 23, 2023.

[4] Retno, S., Rosnita, L., Anshari, S.F. (2023). Sistem Informasi Pelayanan Cuti Berbasis Web Pada PT Pupuk Iskandar Muda Menggunakan PHP dan MySQL. *TECHSI-Jurnal Teknik Informatika*, 14(1), 33-41.

[5] R. Sagita, U. Enri, and A. Primajaya. (2020) "Klasifikasi Berita Clickbait Menggunakan K-Nearest Neighbor (KNN)," *JOINS (Journal Inf. Syst.)*, vol. 5, no. 2, pp. 230–239.

[6] Mustakim, G. O. F. (2016). Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa, 13(2), 195–202.

[7] Retno, S., Dinata, R.K., Hasdyna, N. (2023). Evaluasi model data chatbot dalam natural language processing menggunakan k-nearest neighbor. *Jurnal CoSciTech (Computer Science and Information Technology)*. 4(1): 146-153.

[8] I. Jaya, A. Hizriadi, and E. S. Purba. (2018). "Klasifikasi Surat Laporan Kehilangan Kepolisian Menggunakan Algoritma K – Nearest Neighbor," *TECHSI - J. Tek. Inform.*, vol. 10, no. 2, p. 120.

[9] A. Asrianda, R. Risawandi, and G. Gunarwan. (2019). "Determining Lectural Evaluation in Faculty of Engineering Malikussaleh University Using K-NN," *TECHSI - J. Tek. Inform.*, vol. 11, no. 2, p. 307.

[10] Tang, Y., Jing, L., Li, H., & Atkinson, P. M. (2016). A multiple-point spatially weighted k-NN method for object-based classification. *International Journal of Applied Earth Observation and Geoinformation*, 52, 263–274. <https://doi.org/10.1016/j.jag.2016.06.017>