

Implementation of TF-IDF and XGBoost Algorithms in Scientific Paper Classification

Sany^{1*}, Robet², Joni³

^{1,2} Informatics Engineering, STMIK Time, Medan, Indonesia

³ Information System, STMIK Time, Medan, Indonesia

Sanytn93@gmail.com^{1*}, Robertdetime@gmail.com², Joni.hgw@gmail.com³

Abstract

The rapid growth of scientific publications in the field of informatics demands an accurate and efficient automated classification system. This study aims to implement TF-IDF as a feature extraction method and XGBoost as a classification model to categorize scientific papers, particularly based on their titles. The dataset consists of 1,000 scientific paper titles in the field of informatics, which were collected and processed using text preprocessing techniques. The XGBoost model was trained using TF-IDF vector representations, and hyperparameter tuning was applied to enhance model performance, focusing on parameters such as `learning_rate`, `max_depth`, and `n_estimators`. Evaluation results show that the developed system achieved an accuracy of 81%, with solid performance in distinguishing between “Computer Science” and “Non-Computer Science” categories. This study demonstrates that the combination of TF-IDF and XGBoost is effective for short-text classification such as scientific titles and has potential for further development in multi-class classification and more complex datasets.

Keywords: Text Classification, TF-IDF, XGBoost, Scientific Paper, Hyperparameter Tuning.

1. Introduction

In today’s digital era, the volume of scientific publications in the field of informatics engineering continues to grow rapidly. With various topics and subtopics covering algorithms, computer architecture, cybersecurity, and software development, the management and classification of scientific works have become complex yet crucial tasks. An effective classification process can assist researchers in finding relevant literature and support efforts in innovation as well as the advancement of knowledge. In the academic world, scientific works are often classified into various disciplines, including Computer Science, Physics, Mathematics, Statistics, Quantitative Biology, and Quantitative Finance. Each discipline has its own focus and methodology, reflecting the diversity of research across different fields [1]. In this study, particular attention is given to the classification and prediction of scientific works in the field of informatics engineering. By directing the focus toward informatics engineering, this research aims to develop effective methods for managing and grouping relevant scientific literature, thereby supporting researchers in finding specific and up-to-date information in the field of text data classification [2].

Machine Learning–based models that can be used for text classification include SVM, CNN, and XGBoost. SVM (Support Vector Machine) is a machine learning algorithm applied to classification and regression tasks [3]. CNN (Convolutional Neural Network) involves the process of identifying and adjusting optimal parameter values in its architecture to improve model performance [4]. XGBoost models and text representation techniques such as TF-IDF (Term Frequency–Inverse Document Frequency) have been proven effective in text classification tasks [5]. XGBoost is an efficient and powerful boosting algorithm for classification and regression problems. TF-IDF is a statistical technique used to assess the importance of words in a document relative to a collection of documents. By combining the two, it is possible to create a model capable of capturing essential text features and performing accurate classification [6]. This research focuses on the implementation of these two techniques for the classification of scientific works. By understanding and effectively applying TF-IDF and XGBoost, it is expected that classification performance can be improved, thereby facilitating the management and retrieval of the ever-growing body of scientific information.

2. Theoretical Basis

2.1. Scientific paper

A scientific paper is one of the primary forms of communication in the academic and research world. This document contains research findings, ideas, or systematic studies that are structured based on scientific methods. Through a scientific paper, authors can present new information, solve problems, or discuss specific topics in depth within the relevant field of study.

2.2. Text classification

Text classification refers to the process of categorizing textual documents into predefined classes by analyzing and leveraging the inherent features or patterns within the text.

2.3. TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) is a statistical weighting scheme commonly employed to measure the significance of a term within a document in relation to a larger collection of documents (corpus). This method assigns higher weights to terms that appear frequently in a specific document but less frequently across the entire corpus, thereby capturing their discriminative power. Owing to its effectiveness in identifying salient textual features, TF-IDF has become one of the most widely used techniques in Natural Language Processing (NLP), particularly in tasks such as information retrieval, text mining, and document classification.

2.4. XGBoost

XGBoost is an ensemble-based machine learning algorithm used to address both classification and regression problems. This algorithm combines multiple weak learners to improve prediction accuracy. Boosting is an ensemble technique in which weak models are trained sequentially to correct the errors made by their predecessors. The final result is a combination of all models that produces more accurate predictions.

3. Analysis

Problem analysis is the process of breaking down a problem into smaller components to be studied, with the aim of facilitating a better understanding of the issues within an information system. The outcome of system analysis is a solution to the problems identified in the specification of the new system.

The system analysis stage is a critical step, in which the approach involves identifying problems within the currently operating system and evaluating each existing work procedure. In this way, it is possible to gain an understanding of the existing issues, the challenges faced by the running system, their impacts, and the aspects that need to be considered in validating the objectives of the system being designed before improvements are implemented.

The requirements analysis of the proposed system encompasses two main aspects: functional requirements analysis and non-functional requirements analysis. Functional software analysis can be explained using use cases as illustrated in Figure 3.3 below:

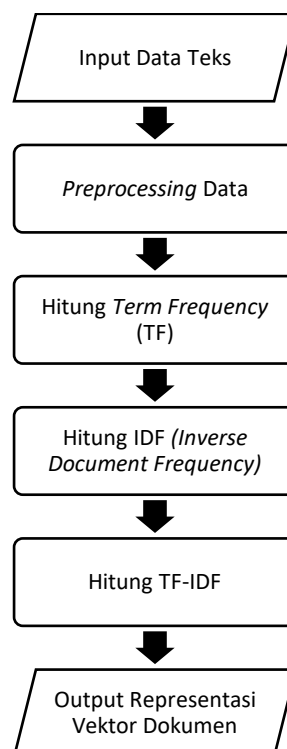


Fig. 1: TF-IDF Process Flow Diagram.

4. A Results and Discussion

4.1. Results

The results of the researcher's study are shown in the following explanation

1. Sign-in Page

Allows users to enter their email and password to access the system. There is also an option to navigate to the Sign-up page for new users. The design of the Sign in page can be seen in the image below:

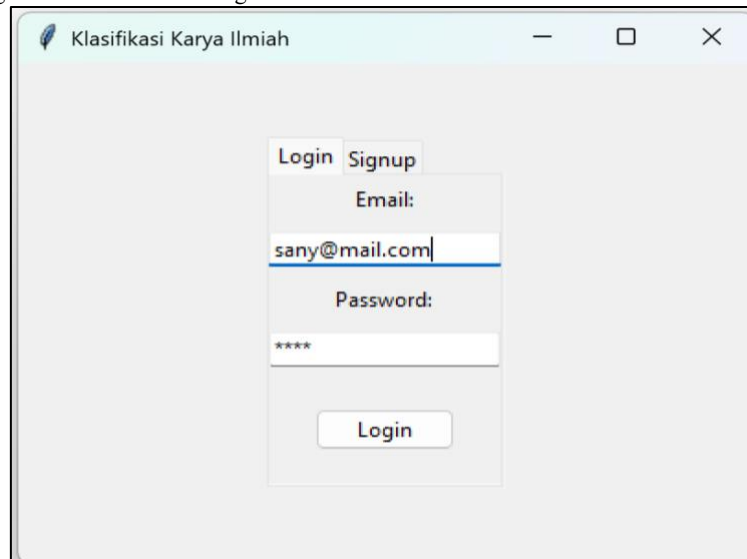


Fig. 2: Sign in Page Design.

2. Home Page

This page serves as the main navigation center after the user successfully logs into the system. The design of the Home page can be seen in the image below:

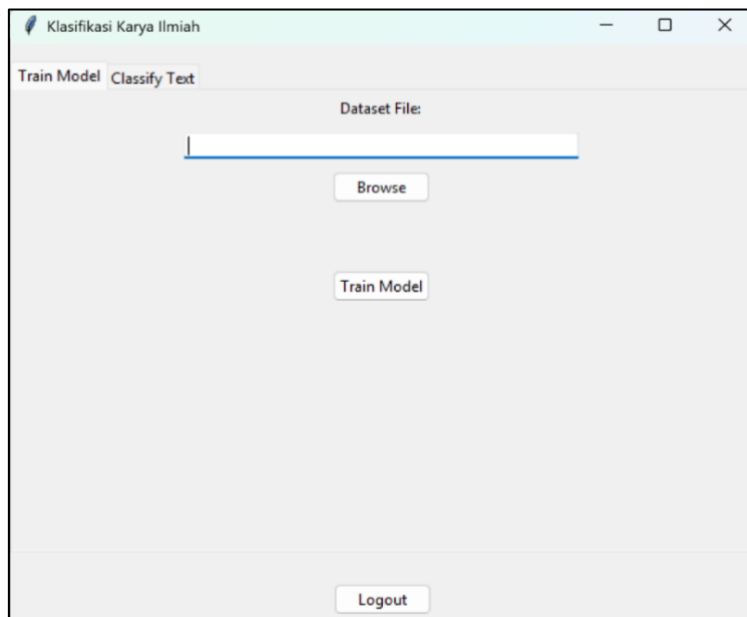


Fig. 3: Home Page Design.

3. Scientific Paper Classification Page

This page provides an input field for entering the title of the scientific paper to be classified. The design of the Scientific Paper Classification Page can be seen in the image below:

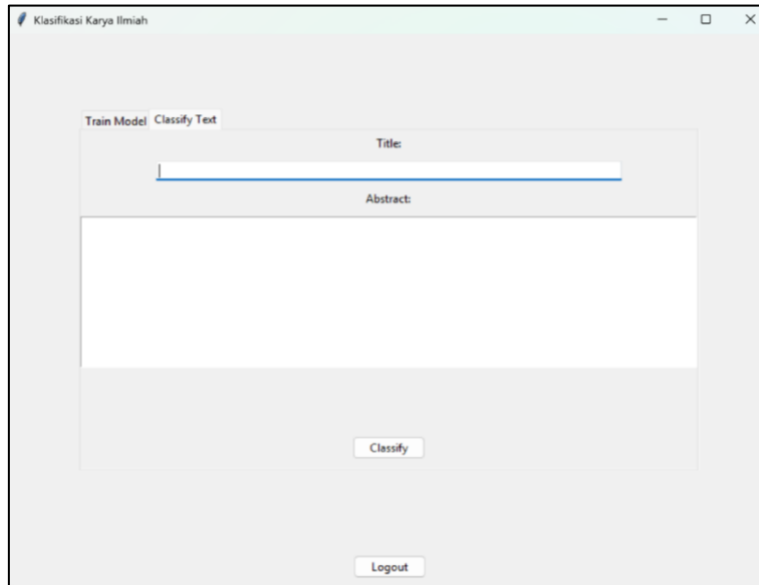


Fig. 4: Scientific Paper Classification Page Design.

4. Model Training Page

On this page, there is a browse option for uploading the dataset. The design of the Model Training Page can be seen in the image below:

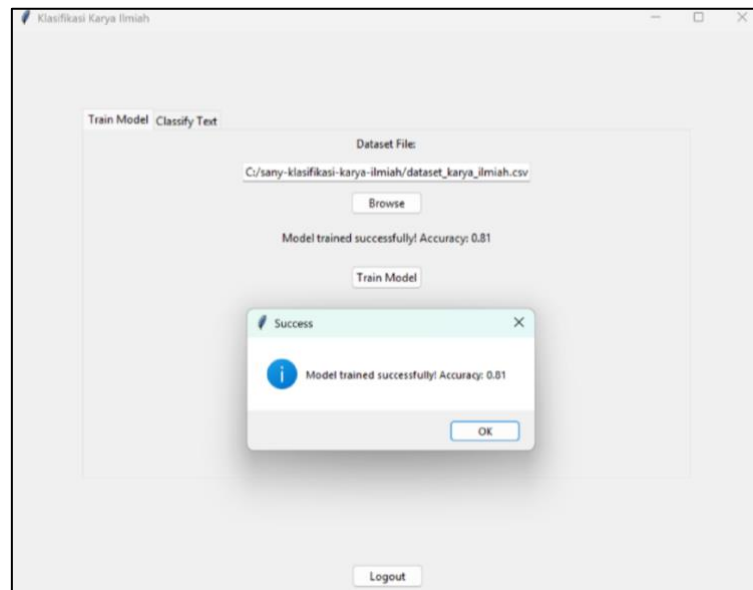


Fig. 5: Model Training Page Design.

4.2. Discussion

The classification results show that the XGBoost algorithm is capable of delivering good performance with an accuracy of 81%. Furthermore, the model demonstrates better performance on class 0 (Non-Computer Science) compared to class 1 (Computer Science), as indicated by the higher recall value.

The advantages of applying TF-IDF and XGBoost:

1. Efficiency of Text Processing

TF-IDF is a simple yet effective feature extraction method for identifying important words, enabling the transformation of text into numerical representations to be carried out quickly and efficiently.

2. Competitive Model Performance

XGBoost is known as a highly powerful boosting algorithm, and in this study, it has been shown to produce good classification results with an accuracy above 80% without requiring high computational infrastructure.

3. Ease of Implementation and Interpretation

This combination is relatively easy to implement, and the classification results can be logically explained based on word weights and decision trees, making it suitable for applications in academic environments.

The shortcomings and limitations include:

1. Data Imbalance

The model tends to perform better on the majority class (Non-Computer Science), while recall on the minority class still needs to be improved using balancing methods such as SMOTE or class weight adjustment.

2. Scalability to Multi-Category Classification

The current model is limited to binary classification. To extend it to multiple categories or subfields of computer science, further development is required, both in preprocessing and in the structure of the classification model.

5. Conclusion

Based on the results of the study on the classification of scientific papers using the XGBoost algorithm and TF-IDF features, it can be concluded that the model demonstrates strong performance with an accuracy of 81%. The model shows high precision in classifying scientific papers into two classes: Computer Science and Non-Computer Science.

From the evaluation results, it was observed that the model performs better in classifying the Non-Computer Science class compared to the Computer Science class. This finding indicates that the distribution of data and the presence of word features in the text significantly influence model performance.

Overall, the XGBoost approach combined with TF-IDF-based feature extraction has been proven effective for large-scale academic text classification tasks.

References

- [1] A. C. Nisha, G. I. Marthasari, and G. W. Wicaksono, "Klasifikasi Abstrak Jurnal Repositor di Teknik Informatika UMM Menggunakan Metode Neighbor Weighted K-Nearest Neighbor," *Jurnal Repositor*, vol. 3, no. 3, 2021.
- [2] R. Nuraeni, A. Sudiarjo, and R. Rizal, "Perbandingan Algoritma Naïve Bayes Classifier dan Algoritma Decision Tree untuk Analisa Sistem Klasifikasi Judul Skripsi," *Innovation in Research of Informatics (INNOVATICS)*, vol. 3, no. 1, 2021.
- [3] M. I. Maulana, K. M. Lhaksana, and M. Dwifebri, "Klasifikasi Komentar Toxic Pada Sosial Media Menggunakan SVM, Information Gain dan TF-IDF," *eProceedings of Engineering*, vol. 10, no. 5, 2023.
- [4] R. Hayami and S. Mohnica, "Klasifikasi multilabel komentar toxic pada sosial media twitter menggunakan convolutional neural network (CNN)," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 4, no. 1, pp. 1–6, 2023.
- [5] D. Safitri and T. A. Fitri, "Perbandingan Algoritma XGBoost dan SVM Dalam Analisis Opini Publik Pemilihan Presiden 2024," *Indonesian Journal of Computer Science*, vol. 13, no. 3, 2024.
- [6] Nurdin, N., Suhendri, M., Afrilia, Y., & Rizal, R. (2021). Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (NBC). *SISTEMASI: Jurnal Sistem Informasi*, 10(2), 268–279.
- [7] Nurhadi, A. (2012). Implementasi algoritma naive bayes classifier berbasis particle swarm optimization (PSO) untuk klasifikasi konten berita digital bahasa Indonesia. *Sari*, 2(3), 48–56.
- [8] Krisnandi, D., Ambarwati, R. N., Asih, A. Y., Ardiansyah, A., & Pardede, H. F. (2023). Analisis Komentar Cyberbullying Terhadap Kata Yang Mengandung Toksisitas Dan Agresi Menggunakan Bag of Words dan TF-IDF Dengan Klasifikasi SVM. *Jurnal Linguistik Komputasional*, 6(2), 36–41.
- [9] Sari, A. K., Irsyad, A., Aini, D. N., & Ginting, S. E. (2024). Analisis Sentimen Twitter Menggunakan Machine Learning untuk Identifikasi Konten Negatif. *Adopsi Teknologi Dan Sistem Informasi (ATASI)*, 3(1), 64–73.
- [10] Prameswari, M., Kania, P. E., De Ayu, I. G., & Harnoko, S. N. P. (2024). Penerapan Metode Stacking Ensemble Untuk Klasifikasi Status Pinjaman Nasabah Bank. *PROSIDING SEMINAR NASIONAL SAINS DATA*, 4(1), 802–811.
- [11] Warda, F., Fajri, F. N., & Tholib, A. (2023). Classification of Final Project Titles Using Bidirectional Long Short Term Memory at the Faculty of Engineering Nurul Jadid University. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, 12(3), 356–362.
- [12] B. Y. Geni, D. Ramayanti, and A. Ratnasari, "IMPLEMENTASI SISTEM POIN OF SALE TERINTEGRASI BERBASIS PYTHON," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 4, pp. 4387–4393, 2024.
- [13] Gumanti, & Elanda, A. (2022). Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Topik Skripsi Mahasiswa di Fakultas Ilmu Komputer. Doctoral Dissertation, Universitas Lancang Kuning.
- [14] Wijayaningrum, V. N., & Lestari, V. A. (2022, September). Jupyter lab platform-based interactive learning. In *2022 International Conference on Electrical and Information Technology (IEIT)* (pp. 295-301). IEEE.