Fake News Detection in Indonesian Language Using a Deep Learning Approach with Indo-BERT

Muhammad Rijal[1*], Harun Musa[2], Faula Yuniarta Seli[3], Nur Ilham Asnawi[4], Ahmad Maruf Firman[5]

[1,2,3,4,5]Politeknik Negeri Ujung Pandang, Indonesia

**) Corresponding author: muhammad.rijal@poliupg.ac.id*

Abstract

The development of digital technology and social media has increased people's exposure to information, but it also raises serious challenges in the form of the spread of fake news. This research aims to develop an Indonesian fake news detection system by utilizing the Indo-BERT model with a Deep Learning approach, a transformer model that has been trained using a large corpus of Indonesian language. The research dataset consists of thousands of articles from CNN Indonesia, Kompas, and Tempo as the original news, and TurnBackHoax as the source of fake news. After going through the text pre-processing stage, the Indo-BERT model was fine-tuned for binary classification. The test results showed excellent performance with an accuracy of 92%, F1-score of 0.92, and ROC value of 0.98, confirming the model's ability to consistently distinguish between real news and hoaxes. The trained model is then integrated into a Web-based application, so that it can be used directly by the public to verify news. In addition to making technical contributions in the Natural Language Processing (NLP) domain, this research also emphasizes social and educational dimensions, namely supporting digital literacy, increasing critical awareness, and strengthening technology-based learning strategies in dealing with misinformation.

Keywords: Fake News, Artificial Intellegence, Deep Learning, Indo-BERT

Introduction

The development of digital technology and social media has brought major changes in the way people obtain, disseminate, and consume information. On the one hand, this digital transformation provides faster, wider and more dynamic access to information. However, on the other hand, this convenience also opens up opportunities for the spread of inaccurate or misleading information, known as fake news. This phenomenon is not only happening in Western countries, but is also a serious issue in Indonesia. According to research results, 44.3% of respondents admitted to receiving fake news or misinformation every day (Antony Lee, 2020).

In Indonesia, the existence of fake news is often associated with crucial issues such as elections, public health, and social conflicts. For example, in the political context, fake news can shape biased public opinion and reduce trust in state institutions (Adila et al., 2023). In the health sector, misinformation during the COVID-19 pandemic has an impact on public compliance with health protocols (Marie Derstroff et al., 2023). Meanwhile, in social life, the rise of hoaxes can strengthen polarization between groups and reduce the quality of social interactions (Alanzi, 2023). Thus, fake news detection is not only important in the realm of information technology, but also has major implications for social science and community development (Vinay et al., 2025).

In addition, the issues of digital literacy and education are particularly relevant in this context. Low media literacy skills among Indonesians, including students, make them vulnerable to exposure to fake news (Roshinta et al., 2023). According to a study in in the field of education, digital literacy not only includes the ability to use technology, but also includes critical thinking skills in filtering and evaluating information. Therefore, the development of fake news detection technology can serve multiple functions: as a tool for information verification, and also as an

educational tool that increases people's critical awareness of the truth of information.

Technologically, various approaches have been used in detecting fake news including the use of machine learning algorithms such as Naïve Bayes and Convolutional Neural Networks (Kurnia et al., 2024). The Indonesian language itself is unique in the form of morphological variations, the use of nonstandard words, as well as foreign language mixtures, which add challenges to text processing. In this case, deep learning-based approaches, especially transformer-based models, offer a more adaptive solution (Yodi Prayoga et al., 2021). Indo-BERT, a BERT model trained with a large Indonesian corpus (Nababan et al., 2024), enables better context understanding and more accurate semantic representation than traditional approaches (Mudding, 2024).

This research aims to apply Indo-BERT in Indonesian fake news detection and develop a web-based application that can be used by the wider community. The main focus of this research is on how NLP technology can be utilized not only as a technical system, but also as part of a social and educational solution. By providing an accurate and easy-to-use detection tool, this research is expected to support digital literacy, increase public critical awareness, and strengthen educational efforts in facing the challenges of the digital information age.

Literature Review

Fake News

Fake news is false information that is deliberately created without factual basis, but packaged as if it were true with the aim of misleading the public and often carrying a certain political agenda (Wiladi & Afrianti, 2024). Hoax news generally contains information that is untrue, false, or sourced from parties that are not credible, and its creation can be driven by a variety of objectives such as seeking attention, gaining profits, and shaping certain public opinions, so it is very important for the public to be able to recognize and avoid the spread of this kind of news. To deal with hoaxes and prevent their negative impacts, the government has actually prepared various legal foundations that are quite strong.

Cross-disciplinary studies confirm that fake news cannot be understood only as a technological problem, but rather a social phenomenon related to cognitive psychology, communication behavior, and social network dynamics (Weiss et al., 2021). Highlighting the role of cognitive bias (confirmation bias), social media algorithms, and information consumption patterns in accelerating the diffusion of fake news. This emphasizes the need for an interdisciplinary approach that combines social science with technology to effectively tackle fake news (Majerczak & Strzelecki, 2022).

Digital literacy

Digital literacy is a person's skill in mastering, managing, assessing, and utilizing digital technology in an appropriate, wise, and responsible manner (Syafrial, 2023). This concept is not only limited to technical mastery in operating digital devices such as computers, smart phones, and the internet, but also includes the ability to think critically, apply ethics, awareness of security aspects, and understanding of the social and cultural influences arising from the use of digital technology (Ririen & Daryanes, 2022).

Media and Information Literacy as a core competency for digital citizens, includes the ability to access, evaluate, and use information critically (Teuku et al., 2023). In Indonesia, low digital literacy is still one of the factors of vulnerability to hoaxes, especially among students. Educational interventions based on inoculation theory such as the Bad News Game have been shown to increase psychological resistance to information manipulation (Rahmanto et al., 2023). In this context, the integration of artificial intelligence (AI) technology into the digital literacy curriculum can be an innovative strategy to foster critical thinking skills, information verification habits, and a deeper understanding of how algorithms and media work. Thus, the integration of AI technology in the digital literacy curriculum can be a strategy to foster critical thinking and verification habits in learners.

Artificial Intelligence

Artificial Intelligence (AI) is a field of computer science that focuses on developing systems that can mimic human cognitive abilities such as learning, reasoning and decision-making. AI has developed rapidly in the last two decades with wide applications in various fields, ranging from recommendation systems, facial recognition, health data analysis, to fake news detection (Chu-Ke & Dong, 2024). In the context of this research, AI is positioned not only as a technical tool to perform automatic classification, but also as a social instrument that supports digital literacy and critical awareness. Thus, AI has a dual role, accelerating the data analysis process while providing a significant social impact in mitigating the spread of misinformation (Santos, 2023).

Advances in Artificial Intelligence (AI) have brought significant transformations in the field of Natural Language Processing (NLP). One important milestone is the arrival of Deep Learning, an approach based on deep neural networks that is capable of extracting data representations hierarchically and automatically. Unlike classical machine learning methods that rely on manual feature engineering, Deep Learning can learn complex patterns in data end-to-end, making it more adaptive in handling natural language diversity (Rofia Abada et al., 2022).

Deep Learning

Deep Learning is a branch of machine learning that uses artificial neural networks architecture with many hidden layers to extract complex data representations (Chu-Ke & Dong, 2024). The main advantage of deep learning is its ability to perform automatic feature learning, making it very effective for handling large amounts of data such as text, images, and sound (Aïmeur et al., 2023). In the natural language processing domain, deep learning has supported the development of transformer models such as BERT, which significantly improves the performance on various NLP tasks including text classification. Therefore, this research utilizes a deep learning approach to fine-tune the Indo-BERT model, so as to identify fake news with high accuracy on Indonesian text (Arora, 2024).

NLP

Natural Language Processing (NLP) is a branch of AI that focuses on the interaction between computers and human language, specifically how machines can understand, process, and generate natural language (Pan et al., 2023). NLP techniques are used for various applications, such as machine translation, chatbots, sentiment analysis, and fake news detection. In recent years, the development of transformer-based models such as BERT, GPT, and their variants has brought great leaps in NLP accuracy and efficiency. For Indonesian, the Indo-BERT model comes as an adaptation of BERT trained with a large-scale Indonesian corpus, making it more contextualized in understanding sentence structure, vocabulary, and linguistic variations unique to Indonesia. In this research, NLP becomes the main foundation in building a hoax detection system that can work effectively on Indonesian news texts (Kaushik, 2023).

Indo-BERT

Recent research has shown that Indonesian language pre-training models such as IndoBERT and its lightweight variant (IndoBERT-lite) produce state-of-the-art performance on various NLP tasks. Benchmarks such as IndoNLU and IndoLEM prove that transformer-based language representations are far superior in understanding the complexity of Indonesian compared to previous approaches (Tobing et al., 2025).

In addition, the development of domainspecific models such as IndoBERTweet shows that vocabulary adaptation for the social media domain can improve the performance of informal text classification. This fact is relevant because most hoaxes circulate through social media platforms (Koto et al., 2021).

Method

Research Framework

The stages of this research began with data collection in the form of online news and social media posts which were then verified through fact-checking agencies. Next, text preprocessing is carried out, including normalization of letters, removal of punctuation marks and stopwords, and tokenization to match the model input format. The processed data is then used in the Indo-BERT fine-tuning stage, which adjusts the pre-trained model for the binary classification task between real and fake news. The final stage is implementation into a Flask-based application, where the trained model is integrated into a web system so that it can be used by users to verify the veracity of news directly.

Dataset

The dataset used in this research comes from the Indonesian Fact and Hoax Political News Dataset on Kaggle, which contains thousands of Indonesian news articles labeled as real news or fake news. Real news data is obtained from three national news portals, namely CNN Indonesia as many as 7,709 articles, Kompas 4,750 articles, and Tempo 6,501 articles, while fake news data is taken from TurnBackHoax with a total of 10,381 articles that have been verified as hoaxes.

1. Indo-BERT Model

In this study, the model used is the pretrained IndoBERT Base model with FineTuning with dense layers and softmax using Hyperparameter. In testing using evaluation metrics such as accuracy, precision, Recall, F1score.

2. Application Implementation

The trained model is implemented into a Flask-based application as a user interface. The mechanism is simple: the user enters the news text, the system then processes the input through the NLP pipeline, and the Indo-BERT model produces a classification with two main categories, namely Fake or Real, along with a probability score to indicate the confidence level of the model. With this design, the application is not only practical and accessible, but can also be used directly in the context of research and public use.

3. Social Integration and Education

Beyond its technical function as a classification tool, the app is also designed with a social and educational dimension. Each prediction result displayed not only serves as a final label, but also as educational feedback for users. The main goal is to support digital literacy programs by raising awareness of the importance of information verification and strengthening critical thinking skills in consuming news. Thus, this application can act as a learning tool as well as a social intervention to build a healthier information ecosystem.

Result and Discussion

Indo-BERT Model Testing Results

a. Confusion Matriks

Confusion matrix shows the distribution of model predictions on test data. Of the 345 Real news, 335 were correctly classified, while only 10 were incorrectly categorized as Fake. Of the 354 Fake news, 311 were correctly detected, while 43 were incorrectly classified as Real. The overall accuracy value reached 92%, with an F1-score of 0.92 as well, indicating a balanced performance between precision and recall. This analysis shows that the model is more likely to be wrong in the case of false negatives (fake news detected as real news).

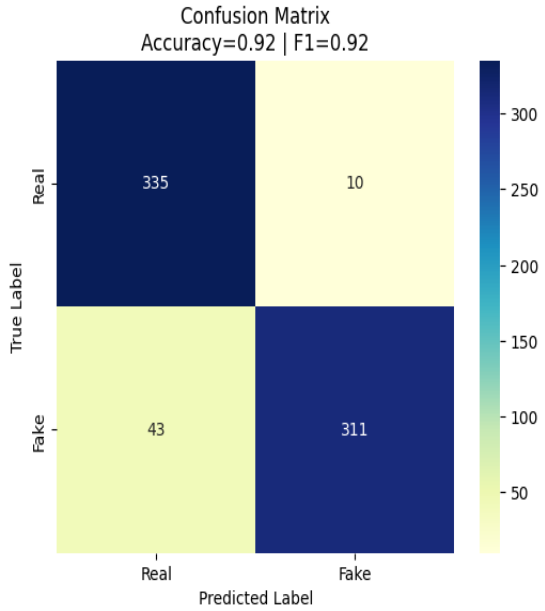

Figure 1. Confusion Matrix

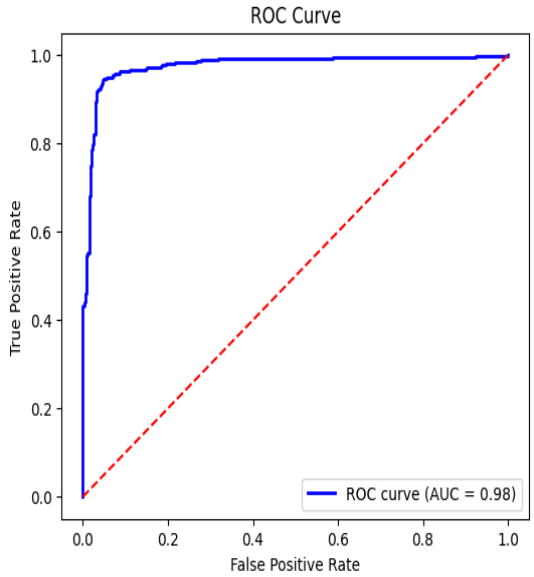

Figure 2. ROC Curve

b. Classification Report

The classification report results show that the IndoBERT model performs very well with an accuracy of 92% and an average F1-score of 0.92. In the Real class, the model achieved a precision of 0.89 and recall of 0.97, which means that most of the real news was recognized correctly although there were still a small number of false predictions. Meanwhile, in the Fake class, the high precision (0.97) indicates that the model rarely misidentifies genuine news as hoaxes, but the lower recall (0.88) indicates that there are still hoaxes that escape detection as genuine news. Overall, the balanced macro and weighted average values prove that IndoBERT is consistent in both classes, although the biggest challenge remains in reducing false negative cases so that hoaxes are not missed.

c. ROC Curve

Receiver Operating Characteristic (ROC) in figure 2 Curve is used to evaluate the tradeoff between true positive rate and false positive rate. The IndoBERT model curve is well above the random diagonal line (random guess) with an AUC of 0.98, confirming that the model has excellent discriminative ability in distinguishing between real and fake news. The closer to the upper left corner of the graph, the higher the classification quality, and this result shows IndoBERT is able to reach a nearoptimal performance level.

WEB Application Implementation Results

The results of implementing the model to the web and testing the application when users enter a news text containing misinformation. The system built with the IndoBERT model automatically classifies the text as "FAKE (HOAX)" with a confidence level of 99.99%. Visualization of the results is displayed through bar indicators that show the predominance of predictions in the Hoax class over Fake or Non-Hoax.

Figure 3. Hoax detection application

Figure 4 shows the opposite case, where the system is given input in the form of text from official news. The prediction result classifies the news as "NOT FAKE (not a hoax)" with the same high confidence level of 99.99%. The probability graph displays the dominance of the Non-Hoax class over Hoax, demonstrating the model's consistency in detecting valid news. This simple yet informative display allows users to understand the classification results quickly and intuitively.

Figure 4. Application detects Non-Hoax

Relevance to Artificial Intelligence

This research emphasizes the role of artificial intelligence not only as a technical tool in performing automatic classification, but also as a social instrument that provides educative feedback to users. The utilization of the IndoBERT model demonstrates the ability of Natural Language Processing (NLP) to understand the complexity of the Indonesian language, including vocabulary variations, syntactic structures, and diverse semantic contexts. Thus, AI doubles as an analytic technology as well as a tool that strengthens people's digital literacy.

Implikasi dalam Social Sciences

The deep learning-based fake news detection application has important implications in social science studies. This technology can be used to examine the pattern of hoax distribution, people's behavior in consuming information, and the social dynamics formed due to misinformation. The results of the analysis can also support the formulation of more effective public policies in mitigating disinformation, maintaining social stability, and strengthening the quality of democracy.

Implications in Education

In education, this application can be integrated into the digital literacy curriculum as a practicum and discussion media. Its utilization encourages students to hone critical thinking skills, evaluate the credibility of information, and understand the ethics of using digital media. Thus, AI technology serves not only as a technical innovation, but also as a learning strategy that fosters information literacy and hoax resistance.

Conclusion

This research proves that the Deep Learning approach with fine-tuning methods is very effective in detecting fake news in Indonesian. The model is able to achieve 92% accuracy with balanced precision, recall, and F1-score, while showing resilience in the face of linguistic complexity typical of Indonesian language. Evaluation based on confusion matrix, classification report, and ROC curve confirms that Indo-BERT has a strong and stable discriminative ability in distinguishing genuine and fake news. The implementation of the model into a Flask-based web application further strengthens the practical value of this research, making it an easily accessible information verification tool for the public. Furthermore, this research highlights the dual role of artificial intelligence, namely as a technical solution as well as a socio-educative instrument that contributes to strengthening digital literacy and critical thinking skills. Thus, Indo-BERT-based fake news detection can be viewed as a strategic innovation that not only answers technological needs, but also

has a significant impact in the social and educational context in Indonesia.

In addition to the technical contribution, this research emphasizes the social and educational implications of applying artificial intelligence in combating misinformation. The developed application not only serves as a verification tool, but also provides educational feedback that encourages critical thinking and supports digital literacy programs. Thus, AIbased fake news detection can be positioned as a technological solution as well as a socialeducative innovation that contributes to mitigating the spread of misinformation in Indonesia.

References

Adila, I., Afnan, N., & Armantari, G. A. (2023). Pengecekan Berita Palsu untuk Mengukur Hoaks di Era Disrupsi Informasi. *Tuturlogi*, *2*(3), 219. https://doi.org/10.21776/ub.tuturlogi.2021.002.03.5

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, *13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Alanzi, T. (2023). Public Perceptions Towards Online Health Information: A Mixed-Method Study in Eastern Province of Saudi Arabia. *Journal of Healthcare Leadership*, *Volume 15*, 259–272. https://doi.org/10.2147/JHL.S431362

Antony Lee. (2020). *Online Hoaxes, Existential Threat, And Internet Shutdown: A Case Study of Securitization Dynamics in Indonesia 1*.

Arora, N. (2024). Combating Generation Of Misinformation Using Ai Enabled Deepfake Technology. *International Journal of Engineering Applied Sciences and Technology*, *9*(08), 17–21. https://doi.org/10.33564/IJEAST.2024.v09i08.004

Chu-Ke, C., & Dong, Y. (2024a). Misinformation and Literacies in the Era of Generative Artificial Intelligence: A Brief Overview and a Call for Future Research. *Emerging Media*, *2*(1), 70–85. https://doi.org/10.1177/27523543241240285

Chu-Ke, C., & Dong, Y. (2024b). Misinformation and Literacies in the Era of Generative Artificial Intelligence: A Brief Overview and a Call for Future Research. *Emerging Media*, *2*(1), 70–85. https://doi.org/10.1177/27523543241240285

Kaushik, Dr. P. (2023). Deep Learning and MachineLearning to Diagnose Melanoma. *International Journal of Research in Science and Technology*, *13*(01), 58–72. https://doi.org/10.37648/ijrst.v13i01.008

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10660–10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

Kurnia, Y., Kusuma, E. D., Kusuma, L. W., Suwitno, & Apridius, W. (2024). Perbandingan Naïve Bayes dan CNN yang Dioptimasi PSO pada Identifikasi Berita Hoax Politik Indonesia. *Bit-Tech*, *6*(3), 340–352. https://doi.org/10.32877/bt.v6i3.1225

Majerczak, P., & Strzelecki, A. (2022). Trust, Media Credibility, Social Ties, and the Intention to Share towards Information Verification in an Age of Fake News. *Behavioral Sciences*, *12*(2), 51. https://doi.org/10.3390/bs12020051

Marie Derstroff, Victoria E. Härtling, Wilhelmiina Hölttä, Mike H. Traub, Linda A.P.J. van der Linden, & James C. Thomas. (2023). Stemming the Tide of Disinformation in Public Health. *South Eastern European Journal of Public Health*, 132–146. https://doi.org/10.56801/seejph.vi.374

Mudding, A. A. (2024). Mengungkap Opini Publik: Pendekatan BERT-based-caused untuk Analisis Sentimen pada Komentar Film. *Journal of System and Computer Engineering (JSCE)*, *5*(1), 36–43. https://doi.org/10.61628/jsce.v5i1.1060

Nababan, W. R., Rahmadani, N., Tamba, W. O. V., & Hidayat Nst, T. K. (2024). Tantangan Bahasa di Era Digital Terhadap Kesalahan Berbahasa Dalam Komunikasi Media Sosial. *Jurnal Bahasa Daerah Indonesia*, *1*(3). https://doi.org/10.47134/jbdi.v1i3.2602

Pan, Q., Zhou, J., Yang, D., Shi, D., Wang, D., Chen, X., & Liu, J. (2023). Mapping Knowledge Domain Analysis in Deep Learning Research of Global Education. *Sustainability*, *15*(4), 3097. https://doi.org/10.3390/su15043097

Rahmanto, A. N., Yuliarti, M. S., & Naini, A. M. I. (2023). Fact Checking dan Digital Hygiene: Penguatan Literasi Digital sebagai Upaya Mewujudkan Masyarakat Cerdas Anti Hoaks. *PARAHITA : Jurnal Pengabdian Kepada*

Masyarakat, *3*(2), 77–85. https://doi.org/10.25008/parahita.v3i2.85

Ririen, D., & Daryanes, F. (2022). Analisis Literasi Digital Mahasiswa. *Research and Development Journal of Education*, 8(1), 210. https://doi.org/10.30998/rdje.v8i1.11738

Rofia Abada, Abdulhalim Musa Abubakar, & Muhammad Tayyab Bilal. (2022). An Overview on Deep Leaning Application of Big Data. *Mesopotamian Journal of Big Data*, *2022*, 31–35. https://doi.org/10.58496/MJBD/2022/004

Roshinta, T. A., Kumala, E., & Dinata, I. F. (2023). Sistem Deteksi Berita Hoax Berbahasa Indonesia Bidang Kesehatan. *Remik*, *7*(2), 1167–1173. https://doi.org/10.33395/remik.v7i2.12369

Santos, F. C. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. *Journalism and Media*, *4*(2), 679–687. https://doi.org/10.3390/journalmedia4020043

Syafrial, H. (2023). *Literasi Digital*. PT Nas Media Indonesia.

Teuku, Z., Akmal, S., Putri, N., & Maulida, T. (2023). Pengabdian Literasi Digital bagi Siswa Pesantren Aliyah Di Banda Aceh dan Aceh Besar. *Jurnal Pengabdian Multidisiplin*, *3*(2). https://doi.org/10.51214/japamul.v3i2.671

Tobing, C. J. L., IGN Lanang Wijayakusuma, & Luh Putu Ida Harini. (2025). Perbandingan Kinerja IndoBERT dan MBERT Untuk Deteksi Berita Hoaks Politik dalam Bahasa Indonesia. *JST (Jurnal Sains Dan Teknologi)*, *14*(1), 114–123. https://doi.org/10.23887/jstundiksha.v14i1.92126

Vinay, R., Spitale, G., Biller-Andorno, N., & Germani, F. (2025). Emotional prompting amplifies disinformation generation in AI large language models. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1543603

Weiss, A. P., Alwan, A., Garcia, E. P., & Kirakosian, A. T. (2021). Toward a Comprehensive Model of Fake News: A New Approach to Examine the Creation and Sharing of False Information. *Societies*, *11*(3), 82. https://doi.org/10.3390/soc11030082

Wiladi, G. J. |, & Afrianti, D. M. (2024). Pengaruh Literasi Media Digital Terhadap Tindakan Penyebaran Berita Palsu Pada Mahasiswa Universitas Bhayangkara. *Jurnal Ilmiah Wahana Pendidikan*, *10*(21), 352–360. https://doi.org/https://doi.org/10.5281/zenodo.14405369

Yodi Prayoga, A., Id Hadiana, A., & Rakhmat Umbara, F. (2021). Deteksi Hoax pada Berita Online Bahasa Inggris Menggunakan Bernoulli Naïve Bayes dengan Ekstraksi Fitur Tf-Idf. *Jurnal Syntax Admiration*, *2*(10), 1808–1823. https://doi.org/10.46799/jsa.v2i10.327