Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

CLICKBAIT DETECTION IN INDONESIA HEADLINE NEWS USING INDOBERT AND ROBERTA

Muhammad Edo Syahputra-[1]*, Ade Putere Kemala-[2], Dimas Ramdhan-[3]

Computer Science Department[1*), 2], Data Science / School of Computer Science[3]
Bina Nusantara University
Jakarta, Indonesia
muhammad.syahputra002@binus.ac.id[1] , ade.kemala@binus.ac.id[2], dimas.ramdhan@binus.ac.id[3]

(*) Corresponding Author

Abstract

This paper explores clickbait detection using Transformer models, specifically IndoBERT and RoBERTa. The objective is to leverage the models specifically for clickbait detection accuracy by employing balancing and augmentation techniques on the dataset. The research demonstrates the benefit of balancing techniques in improving model performance. Additionally, data augmentation techniques also improved the performance of RoBERTa. However, it resulted differently for IndoBERT with slightly decreased performance. These findings underline the importance of considering model selection and dataset characteristics when applying augmentation. Based on the result, IndoBERT, with a balanced distribution, outperformed the previous study and the other models used in this research. This study used three dataset distribution settings: unbalanced, balanced, and augmented with 8513, 6632, and 15503 total data counts, respectively. Furthermore, by incorporating balancing and augmentation techniques, the research surpasses previous studies, contributing to the advancement of clickbait detection accuracy, contributing to the advancement of clickbait detection accuracy with 95% accuracy in f1-score with unbalanced distribution. However, the augmentation method in this study only improved the RoBERTa model. Moreover, performance might be boosted by gathering more varied datasets. This work highlights the value of leveraging pre-trained Transformer models and specific dataset-handling techniques. The implications include the necessity of dataset balancing for accurate detection and the varying impact of augmentation on different models. These insights aid researchers and practitioners in making informed decisions for clickbait detection tasks, benefiting content moderation, online user experience, and information reliability. The study emphasizes the significance of utilizing state-of-the-art models and tailored approaches to improve clickbait detection performance.

Keywords: Clickbait Detection; Transformer; Deep Learning; Data Augmentation

Abstrak

Makalah ini mengeksplorasi pendeteksian clickbait menggunakan model Transformer, khususnya IndoBERT dan RoBERTa. Tujuannya adalah untuk memanfaatkan model khusus untuk akurasi deteksi clickbait dengan menggunakan teknik penyeimbangan dan augmentasi pada dataset. Penelitian menunjukkan manfaat teknik penyeimbangan dalam meningkatkan kinerja model. Selain itu, teknik augmentasi data juga meningkatkan kinerja RoBERTa. Namun hasil yang didapatkan berbeda bagi IndoBERT dengan sedikit penurunan kinerja. Temuan ini menggarisbawahi pentingnya mempertimbangkan pemilihan model dan karakteristik dataset saat menerapkan augmentasi. Berdasarkan hasil tersebut, IndoBERT dengan distribusi berimbang mengungguli penelitian sebelumnya serta model lain yang digunakan dalam penelitian ini. Penelitian ini menggunakan tiga skema dataset untuk melakukan eksperimen yaitu distribusi unbalanced sebanyak 8513 total data, balanced 6631 data, dan augmented data dengan total 15503 data. Selain itu, dengan menggabungkan teknik penyeimbangan dan augmentasi, penelitian ini melampaui penelitian sebelumnya, dengan mendapatkan akurasi sebesar 95% pada model IndoBERT yang divalidasi menggunakan metode f1-score. Namun, berdasarkan eksperimen metode augmentasi tidak memberikan kenaikan pada model IndobERT. Sebaliknya metode augmentasi cukup efektif pada model RoBERTa. Selain itu, dengan menambahkan jumlah data yang lebih bercvariasi dapat meningkat performa model secara signifikan. Pekerjaan ini menyoroti nilai memanfaatkan model Transformer yang telah dilatih sebelumnya dan teknik *penanganan kumpulan data tertentu. Implikasinya termasuk perlunya penyeimbangan dataset untuk deteksi yang akurat dan berbagai dampak augmentasi pada model yang berbeda. Wawasan ini membantu peneliti dan praktisi dalam membuat keputusan yang tepat untuk tugas deteksi clickbait, yang menguntungkan moderasi konten, pengalaman pengguna online, dan keandalan informasi. Studi ini menekankan pentingnya memanfaatkan model canggih dan pendekatan yang disesuaikan untuk meningkatkan kinerja deteksi clickbait.*

Kata kunci: Clickbait Detection; Deep Learning; Deep Learning; Data Augmentasi

INTRODUCTION

Clickbait has become a mainstream approach in online media, where the headline is not aligned with the content. Usually, the headline uses catchy or exaggerated words to attract the reader's attention. This approach considerably negatively impacts society (Bondielli & Marcelloni, 2019; Zhou et al., 2022). In a survey (Bondielli & Marcelloni 2019), the clickbait approach potentially leads to a polarized society. The existence of social media with its algorithm can amplify the distribution of information, including news, which could shape society's discourse in a public space (ShuKai et al., 2017). Such phenomenon also appears in the United States, UK, and arguably worldwide social media.

Many studies have been conducted exploring methods to tackle this clickbait problem. An early approach using a Machine Learning classifier is explored (Abbas et al., 2019; Chakraborty et al., n.d.; Manjesh et al., 2018; Potthast et al., 2016). However, the studies mentioned still lack performance in capturing meaning and validating using Fleiss' within the context of headline news (Zheng et al., 2021).

However, recent studies using the Deep Learning approach have shown promising results (Agrawal, n.d.; Kim, 2014; Zhou et al., 2022). A study (Aju et al., 2022) conducted an empirical study comparing the performance of the Machine Learning and Deep Learning approaches. The result shows that BERT provides maximum efficiency by outperforming the Machine Learning method with 10% accuracy. Also, in the study (Oliva et al., 2022), the divergence measures technique in Deep Learning to tackle dataset availability in clickbait detection outperforms the Machine Learning approach in accuracy.

However, there is still a gap to be filled in Indonesian clickbait detection. Since there are still only a few studies focusing on Bahasa Indonesia using the Deep Learning approach. Moreover, the availability of the dataset in Bahasa Indonesia is also one of the gaps in this research and in Deep Learning Research in general. Also, one of the disadvantages of the Deep Learning approach is that it takes more time to train since a large dataset is needed to extract and learn features (Sirusstara et al., 2022a).

Therefore, this research aims to explore and leverage a pre-trained model to build a classifier model for Indonesian Headline news compared to the previous research (Sirusstara et al., 2022a) using the same dataset from (Hadiyat, 2019). In addition, this research utilized the dataset for balancing and augmentation to improve model accuracy, which was not utilized in the previous study.

RESEARCH METHODS

The methodology proposed in this research consists of three phases: obtaining the dataset, pre-processing, and training using deep learning models.

Dataset

This research obtained the dataset from (William & Sari, 2020). The dataset contains more than 15000 Indonesian headline news from 12 news publishers. Every headline is annotated as clickbait and non-clickbait by the 3 annotators method (Fleiss & Cohen, 1973). Based on annotation results, the dataset has several versions. The version used in this research is mainly from the inter-annotator agreement score of 0.42, the highest Fleiss' K score with distribution, as shown in Table 1.

Table 1. Dataset Label Distribution

Class	Total Count
Clickbait	3316
Non-clickbait	5297
Total	**8513**

Pre-processing

Considering the amount of headline news dataset, which is relatively small, this research used augmentation techniques such as EDA (Wei & Zou, n.d.) and Bootstrapping (Stine, 2016) to balance the distribution. Since the Deep Learning technique requires a large amount of data to achieve great results, constructing the dataset will most likely affect the model's performance dramatically. In addition, data augmentation has been proven to be effective and improve performance in many tasks, including computer vision (Perez & Wang, 2017) and speech recognition (Park et al., 2019). Data augmentation is a technique that utilizes the available data to synthesize new similar data. Therefore, it allows researchers to overcome the scarcity of the dataset available. Furthermore, adding more variety to the data also leads to avoiding overfitting.

Bootstrapping

In this research, a simple balancing method is conducted using resample method. In this method, a random resampling is proceeded by generating extra data points to the dataset based on *bootstrapping* procedure (Stine, 2016). Table 2 is shown the difference in dataset distribution after resampling.

Table 2. Before and After Resampling

Class	Total Count
Clickbait	3316
Non-clickbait	3316
Total	**6632**

EDA

Another approach to data augmentation used in this research is EDA (Easy Data Augmentation). In general, EDA consists of four approaches Synonym Replacement (SR), Random Insertion (RI), Random Swap (RW), and Random Deletion (RD). In the first step, SR's function is initialized to randomly selects n number of words and replaces these words randomly based on selected synonyms from IR's function. IR generates synonyms in n times randomly, excluding stop words to be replaced in SR with random positions. Also operated in n times, RS randomly selects two words in a sentence to swap their positions. Finally, RD measures these selected words in a sentence in a probability of p. Table 3 shows the dataset distribution after EDA.

Table 3. Before and After Augmentation

Class	Total Count
Clickbait	9539
Non-clickbait	5964
Total	**15503**

Deep Learning Technique

This paper focuses on deep learning algorithms, specifically using large models from Indonesian BERT (Koto et al., 2020; Wilie et al., 2020) or called IndoBERT. IndoBERT uses the same architecture as the original BERT with different pre-training datasets, which are unlabeled text-corpus. BERT uses a multi-headed attention mechanism as its method (Vaswani et al., n.d.) and has been proven to outperform other methods in many NLP benchmarks. In other words, BERT is currently known as the state-of-the-art in NLP research.

A paper from (Koto et al., 2020) proposed IndoBERT, which heavily pre-trained the model using mostly news datasets such as Kompas, Tempo, Liputan6, Wikipedia, etc. In total, IndoBERT trained over 220M corpus words. IndoBERT trained purely as a masked language model using the Huggingface framework. They also followed the default configuration for BERT-base, which has 12 hidden layers, 12 attention heads, and feed-forward hidden layers of 3,072d. The model used an Adam optimizer and linear scheduler in pre-training. Moreover, the paper also shows the model's high performance in many tasks, such as summarization, sentiment analysis, named entity recognition, etc.

Unlike (Koto et al., 2020), (Wilie et al., 2020) also proposed IndoBERT with a larger and more general pre-training dataset. The model trained over 4 billion corpus Indonesian words from many sources such as Wikipedia, webpage articles, and Twitter. Therefore, this research implemented a model from (Wilie et al., 2020). However, the model from (Koto et al., 2020) is also considered to see the impact of specifically pre-training datasets from news headlines that are dominantly used in the research, which is suitable for this research.

For training both IndoBERT models, this research used Adam optimizer with a learning rate 2e-5 and batch size of 64 and 10 epochs using one GPU NVIDIA RTX 3080Ti. Furthermore, this research added more than 3000 headline news to the dataset and trained on both models with the same configuration. Then, examine the result for further analysis.

RESULTS AND DISCUSSION

Model Comparison

Table 4 shows the experimental results of IndoBERT (Wilie et al., 2020) and RoBERTa model from hugging face by Cahya, in which the model has been trained with the Indonesian dataset.

Overall, IndoBERT architecture trained with a balanced dataset and without augmentation outperformed other models with (95%) accuracy of f1-score. This accuracy also outperformed previous research using XLM-RoBERTa with (91%) accuracy of f1-score.

Table 4. Experiment Results

Model	Train Test Split 8:2		Average Precision	Average Recall	F1-score
	Train	Test			
indobenchmark/indobert-base-p1					
Imbalance	0.99	0.93	0.96	0.86	0.94
Balance	0.9882	0.9396	0.9259	0.9176	0.9508
Imbalance + Augment	0.9964	0.98	0.907	0.87	0.93
Balance + Augment	0.9952	0.936	0.904	0.934	0.94
cahya/roberta-base-indonesian-522M					
Imbalance	0.98	0.92	0.87	0.92	0.93
Balance	0.9839	90.36	0.84	0.9182	0.91
Imbalance + Augment	0.9925	0.7766	0.6935	0.7589	0.8119
Balance + Augment	0.993	0.9164	0.8764	0.9131	0.9308

This study also experimented with the same model as the previous study, adding the balancing and augmentation method to the dataset. The augmentation method combined with a balanced dataset improved the accuracy by 4%. Thus, this study's model performed better than the previous one.

Also, based on the experiment results shows that balanced data distribution could improve the accuracy of both models. Whereas the augmentation method only boosts performance on the RoBERTa model.

Data Distribution Results

Table 4 also shows the experiment results based on the type of dataset presented in the table. According to the results, they are revealing distinct trends for the IndoBERT and RoBERTa. The experiment using a dataset with a balanced distribution exhibited improved accuracy for IndoBERT and RoBERTA. In contrast with the augmentation method, the model only improves RoBERTA. However, the augmentation method only enhances the performance of the RoBERTa model while slightly decreasing the performance of IndoBERT.

The EDA augmentation method dramatically increases dataset distribution, as Table 3 illustrates. This method leverages the Indonesian WordNET, incorporating operations such as synonym replacement, insertion, swap, and deletion, which introduce noise into the training data. Although the EDA method employs simple operations, the unilingual Indonesian WordNet is better suited for our dataset, which consists of Indonesian headline news.

The amount of augmented dataset could be tuned and equated for each augmentation method. However, this paper used the default parameters, thus producing a different amount of augmented dataset.

Comparison with Previous Study

Table 5 compares the results of this study with the best results from the previous study. Table 5 showcases the performance of IndoBERT and RoBERTa models using the EDA augmentation method and the previous research's best method, including RoBERTa and XLM-RoBERTa models without data augmentation. The results indicate that when combined with balanced data distribution and the augmentation method, the models in this study outperformed the previous research in all testing scenarios.

Table 5. Model Comparison with Previous Study

Model	Augmentation Type	Precision	Recall	F1-score
IndoBERT (This study)	normal	0.96	0.86	0.94
IndoBERT (This study)	Eda	0.92	0.91	0.95
RoBERT (Sirusstara et al., 2022b)	normal	0.87	0.874	8738
RoBERTA	eda	0.87	0.91	0.93
xlmRoBERTa (Sirusstara et al., 2022b)	normal	0.91	0.91	0.91

CONCLUSIONS AND SUGGESTIONS

This research employed two architectural models and three dataset settings: imbalance, balance, and augmented. Despite utilizing the EDA augmentation method, IndoBERT demonstrated superior performance with 95% accuracy in the f1-score when the distribution was balanced. However, as observed in previous studies, augmentation improved accuracy in the RoBERTa model without balancing an augmentation method towards the dataset. To further advance clickbait detection, exploring additional deep learning architectures for future research would be beneficial. Furthermore, collecting more diverse datasets could enhance performance, and investigating the parameters in the augmentation models could be another avenue for future investigation.

REFERENCES

Abbas, M., Ali Memon, K., & Aleem Jamali, A. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. *IJCSNS International Journal of Computer Science and Network Security*, *19*(3), 62.

Agrawal, A. (n.d.). *Clickbait Detection using Deep Learning*. Retrieved September 21, 2022, from https://www.reddit.com/r/news

Aju, D., Kumar, K. A., & Lal, A. M. (2022). Exploring News-Feed Credibility using Emerging Machine Learning and Deep Learning Models. *Journal of Engineering Science and Technology Review*, *15*(3), 31–37. https://doi.org/10.25103/JESTR.153.04

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38–55. https://doi.org/10.1016/J.INS.2019.05.035

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (n.d.). *Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media*.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309/ASSET/001316447303300309.FP.PNG_V03

Hadiyat, Y. D. (2019). Clickbait on Indonesia Online Media. *Journal Pekommas*, *4*(1), 1. https://doi.org/10.30818/jpkm.2019.2040101

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1746–1751. https://doi.org/10.3115/V1/D14-1181

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. 757–770. https://doi.org/10.18653/V1/2020.COLING-MAIN.66

Manjesh, S., Kanakagiri, T., Vaishak, P., Chettiar, V., & Shobha, G. (2018). Clickbait Pattern Detection and Classification of News Headlines Using Natural Language Processing. *2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2017*. https://doi.org/10.1109/CSITSS.2017.8447715

Oliva, C., Palacio-Marín, I., Lago-Fernández, L. F., & Arroyo, D. (2022). *Rumor and clickbait detection by combining information divergence measures and deep learning techniques*. 1–6. https://doi.org/10.1145/3538969.3543791

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). *SpecAugment:*

A Simple Data Augmentation Method for Automatic Speech Recognition. https://doi.org/10.21437/Interspeech.2019-2680

Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *Undefined*.

Potthast, M., Köpsel, S., Stein, B., & Hagen, M. (2016). Clickbait Detection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9626*, 810–817. https://doi.org/10.1007/978-3-319-30671-1_72

ShuKai, SlivaAmy, WangSuhang, TangJiliang, & LiuHuan. (2017). Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Sirusstara, J., Alexander, N., Alfarisy, A., Achmad, S., & Sutoyo, R. (2022a). Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa). *2022 3rd International Conference on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future, AiDAS 2022 - Proceedings*, *September*, 248–253. https://doi.org/10.1109/AiDAS56890.2022.9918678

Sirusstara, J., Alexander, N., Alfarisy, A., Achmad, S., & Sutoyo, R. (2022b). Clickbait Headline Detection in Indonesian News Sites using Robustly Optimized BERT Pre-training Approach (RoBERTa). *2022 3rd International Conference on Artificial Intelligence and Data Sciences: Championing Innovations in Artificial Intelligence and Data Sciences for Sustainable Future, AiDAS 2022 - Proceedings*, 248–253. https://doi.org/10.1109/AIDAS56890.2022.9918678

Stine, R. (2016). An Introduction to Bootstrap Methods. *Http://Dx.Doi.Org/10.1177/0049124189018002003*, *18*(2–3), 243–291. https://doi.org/10.1177/0049124189018002003

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (n.d.). *Attention Is All You Need*.

Wei, J., & Zou, K. (n.d.). *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks*. 6382–6388. Retrieved September 23, 2022, from http://github.

Wilie, B., Vincentio, K., Indra Winata, G., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., Purwarianti, A., & Bandung, I. T. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding* (pp. 843–857). https://aclanthology.org/2020.aacl-main.85

William, A., & Sari, Y. (2020). CLICK-ID: A novel dataset for Indonesian clickbait headlines. *Data in Brief*, *32*, 106231. https://doi.org/10.1016/J.DIB.2020.106231

Zheng, J., Yu, K., & Wu, X. (2021). A deep model based on Lure and Similarity for Adaptive Clickbait Detection. *Knowledge-Based Systems*, *214*, 106714. https://doi.org/10.1016/J.KNOSYS.2020.106714

Zhou, M., Xu, W., Zhang, W., & Jiang, Q. (2022). Leverage knowledge graph and GCN for fine-grained-level clickbait detection. *World Wide Web*, *25*(3), 1243–1258. https://doi.org/10.1007/S11280-022-01032-3