Comparative Analysis of BERT Model and DistilBERT Model for Enhanced Clickbait Headline Structure Detection in Indonesian Online News

Rananggana Trustha Dewangga[1*], Budi Prasetiyo[2]

[1,2] *Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia*

*corr-author: dewangga.trustha@students.unnes.ac.id

Abstract - Clickbait uses sensational or misleading headlines to attract readers, which can degrade information quality in online news. This study presents a comparative evaluation of BERT and DistilBERT for detecting clickbait headline structures in the Indonesian language using the CLICK-ID dataset. The approach examines how class imbalance influences performance by training models on multiple dataset variants created through oversampling, undersampling, and data augmentation. Inputs are tokenized with model specific tokenizers and evaluated with accuracy, precision, recall, and F1-score. Confusion matrices are used to interpret error patterns across classes. Experimental results show that DistilBERT trained on an oversampled dataset achieves 94% for accuracy, precision, recall, and F1-score, while BERT on the same oversampled setting reaches 93%. Models trained on unbalanced data yield the lowest recall and F1 for the clickbait class, confirming the adverse effect of skewed distributions. Augmented and undersampled variants produce slightly lower but competitive results in the 92% to 93% range. Error analysis shows that DistilBERT reduces missed clickbait while maintaining a similar level of false positives, producing more balanced behavior across classes. These results outperform prior CLICK-ID studies and highlight the advantage of transformer architectures combined with effective class balancing for Indonesian clickbait detection.

Keywords: clickbait, text classification, natural language processing, deep learning, transformers based model

I. INTRODUCTION

Online news portals remain vital sources of information, yet profit-driven practices have encouraged the proliferation of clickbait. Clickbait consists of misleading, low-quality headlines engineered to maximize engagement, thereby increasing advertising value and publisher revenue [1]. The phenomenon exploits the curiosity gap through suggestive cues that entice clicks but often fails to deliver on the promised content, leaving readers misled or disappointed [2-3]. This practice degrades information quality and contributes to misinformation, underscoring the need for effective methods to identify and filter such content.

Classifying headlines into clickbait and non-clickbait is an effective way to filter misleading content and is widely approached as a text classification task in natural language processing [4-5]. Text classification is widely used in spam detection, sentiment analysis, and topic labeling, with methods ranging from traditional machine learning such as Support Vector Machines and Naive Bayes to modern deep learning [6]. Traditional models often require extensive feature engineering and struggle with large, unstructured text, which limits their ability to capture complex patterns [7]. Deep learning learns rich representations directly from data, scales to large corpora, and models sequential dependencies, leading to stronger performance for headline classification [6-8].

Previous studies implemented deep learning models such as CNN and Bi-LSTM for clickbait headline classification on the CLICK-ID dataset, reporting 88% and 87% accuracy respectively [9]. However, these models face challenges in capturing bidirectional dependencies in word sequences [5]. CNN process sequences in a single direction which limiting their ability to account for subsequent context, while LSTM although better at modeling sequences, are constrained in representing full-sequence context. Bi-LSTM processes data in both directions but still struggles with long sequences and does not learn all combinations of contextual dependencies [10].

A multilingual Bidirectional Encoder Representations from Transformers (BERT) approach combined with undersampling achieved 91% accuracy on CLICK-ID, however their approach did not assess alternative balancing strategies beyond undersampling [11]. A separate work pairing Bi-LSTM with Word2Vec and random oversampling reached 89% on the same

dataset [12]. These results indicate benefits from class balancing, although the studies focused on recurrent architectures and did not compare multiple balancing methods or transformer-based models [26].

These observations indicate while both traditional deep learning and transformer methods have been explored, systematic comparisons of multiple data-balancing techniques with BERT-based models for Indonesian clickbait detection remain limited [9, 11-12]. To address these challenges, this study employs BERT, which encodes tokens bidirectionally to capture context from both directions [5]. BERT's architecture supports long-range dependency modeling and efficient parallel computation for NLP tasks [10, 13]. Computational cost remains a practical constraint, including for smaller BERT variants [14]. Therefore, this study also evaluates DistilBERT, a compressed model trained via knowledge distillation that retains 97% of BERT's language understanding, with a 40% reduction in size and 60% faster inference [14-15].

Building on these gaps, this study combines transformer-based models with multiple data-balancing strategies, namely oversampling, undersampling, and data augmentation, to evaluate their effect on classification performance in the Indonesian language. Prior reports show that balancing can improve text classification outcomes in general settings [16]. By integrating these strategies with BERT and DistilBERT models and assessing precision, recall, F1-score, and accuracy, this study provides a comparative view against prior results and clarifies how balancing choices influence transformer performance for Indonesian clickbait detection.

II. METHOD

In this study, BERT and DistilBERT models were implemented for clickbait classification using the CLICK-ID dataset. Models were developed using different techniques to address data imbalance, achieving optimal results. These models were then evaluated using predefined metrics and compared to previous research. The workflow of this research study is illustrated in Fig. 1.

A. Dataset

This study utilizes the CLICK-ID dataset, which contains 8,613 annotated news headlines [9]. The dataset includes a "title" column with the news headlines, a "label" column indicating whether a headline is clickbait or non-clickbait, and a "label_score" column, which represents this information numerically 1 for clickbait and 0 for non-clickbait. Table I provides sample entries from the dataset.

B. Data Preprocessing

The data preprocessing in this study consists of several key steps, which include handling missing values, removing irrelevant columns, and performing text cleaning. The first step is identifying and addressing missing values, as they can impact model performance. Depending on the dataset, missing values are either replaced with the mean or mode or removed entirely. The next step involves dropping unnecessary columns to improve analysis efficiency. In this process, the "label" column is removed because the dataset already includes a "label_score" column with numeric values each represent the title type. The final step is text cleaning, which involves converting all text to lowercase through case folding, removing URLs, numbers, punctuation, and unnecessary whitespaces to ensure a clean and consistent dataset for model training.

C. Exploratory Data Analysis

Exploratory data analysis aims to better understand the characteristics of the dataset by visualizing data distribution, examining the proportion of clickbait and non-clickbait labels, and analyzing variations in text length. The analysis begins with creating a Wordcloud to visualize the most frequently occurring words in the dataset. Words are displayed in varying sizes based on their frequency, with common stopwords excluded to focus on more relevant terms for label prediction [17].

Next, the dataset is examined to determine the distribution of clickbait and non-clickbait labels, ensuring balance between the classes. Imbalanced data can lead to biased predictions, making it necessary to apply data balancing techniques if needed [18]. Finally, the number of tokens per headline is analyzed to determine the optimal maximum token length as a hyperparameter [13]. This ensures that the model processes headlines efficiently without omitting important information.

D. Balancing Data

The CLICK-ID dataset contains 8,613 annotated headlines consisting of 5,297 non-clickbait and 3,316 clickbait headlines [9]. This distribution indicates a moderate class imbalance that may bias the model toward predicting the majority class [16]. To address this issue, three balancing techniques were applied: oversampling, undersampling, and data augmentation.

In the oversampling technique, samples from the minority class were duplicated until both classes had equal representation, producing a balanced dataset of

10,595 samples [19]. The undersampling technique reduced the number of majority class samples to match the size of the minority class, resulting in a smaller dataset of 6,632 samples but with balanced class distribution [20]. Data augmentation was performed by generating new clickbait samples using a pre-trained BERT model to replace certain words with contextual synonyms, maintaining semantic meaning while introducing variation [21]. This approach produced a balanced dataset similar in size to the oversampled version.

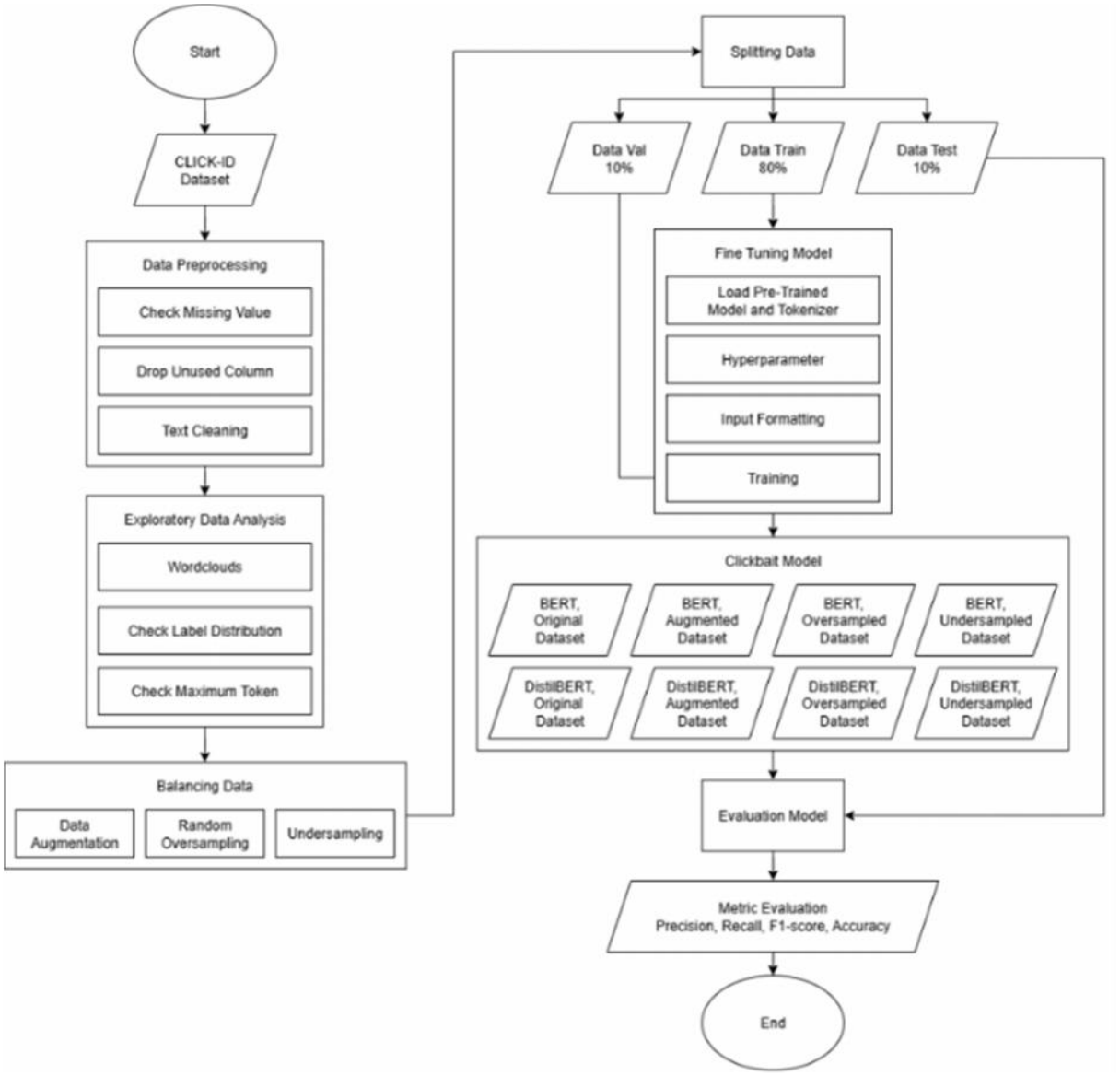

Fig. 1 Research flow

TABLE I
CLICK-ID DATASET SAMPLE

No.	Title	Label	Label_Score
1	BI Kenalkan Standarisasi QR di Kalimantan Selatan	Non-clickbait	0
2	"Disindir Hilang Setelah Masuk Istana, Ini Komentar Teten Masduki hingga Johan Budi"	Clickbait	1
3	"Cepat Kirim Surat Presiden Revisi UU KPK ke DPR, Ini Alasan Jokowi"	Clickbait	1
4	Ribuan SMP di Jateng Jadi Pilot Project Gerakan Menabung Nasional OJK	Non-clickbait	0

Each method offers different trade-offs. Oversampling is simple to implement and restores class balance, but it may lead to overfitting due to duplicated samples [22]. Undersampling can reduce training time and maintain balanced classes, but it may discard potentially useful information from the majority class. [23] Data augmentation adds linguistic variety that can improve generalization, although the quality and relevance of generated text may vary depending on the augmentation process [24].

E. Splitting Data

Data splitting divides the dataset into training, testing, and validation subsets to ensure the model generalizes well. The training set teaches the model patterns, the testing set evaluates performance, and the validation set tunes hyperparameters to prevent overfitting. This study uses an 80:10:10 split, balancing training efficiency and evaluation reliability [6]. Additionally, random splitting minimizes selection bias and enhances generalization.

F. Fine Tuning BERT Model and DistilBERT Model

The fine-tuning process aims to develop and train a model for specific tasks such as clickbait detection, utilizing a pre-trained model as initial knowledge transferred to the new model [25]. In this study, the general model structure was expanded into several variations, which were trained and developed. The details of these model variations are shown in Table II.

After splitting the data, we fine-tuned two pre-trained Indonesian transformer models from Hugging Face, namely bert-base-indonesian-1.5G and distilbert-base-indonesian to leverage rich language representations [26]. Hyperparameters followed established guidance from the BERT literature to balance learning efficiency and overfitting risk [5]. Input text was prepared with the respective tokenizers using special tokens, fixed-length padding or truncation, and attention masks to distinguish real tokens from padding, which supports robust contextual encoding [5, 15]. Training optimized the models to classify headlines into clickbait and non-clickbait with validation monitoring and early stopping. Performance was evaluated using accuracy, precision, recall, and F1-score, and all configurations were summarized in a results table to identify the best performing model.

III. RESULT AND DISCUSSION

This research employs several model variations using BERT and DistilBERT to identify the most effective and optimal model for classifying the structure of clickbait news headlines. Pretrained models from the Hugging Face repository are fine-tuned on multiple dataset variants. To address class imbalance, the experiments use random oversampling, undersampling, and data augmentation.

A. Data Preprocessing

Data preprocessing transform raw data into a clean and structured format for analysis. Missing values are checked to ensure data completeness. Column removal is performed to eliminate unnecessary columns, keeping “Title” and “Label_score” column serves as the numeric class label. Text is then cleaned by converting to lowercase and removing special characters, numbers, URLs, punctuation, and extra spaces using regular expressions. These steps improve data quality and consistency, enabling more reliable training and evaluation. Table III presents a sample of the dataset after data preprocessing.

B. Exploratory Data Analysis

After data preprocessing, Exploratory data analysis is conducted to gain deeper insights into the dataset. In word cloud analysis, the first step is removing stopwords like "and," "in," and "that," as they do not provide meaningful information and may hinder keyword identification. This process results in cleaner text for further analysis. Table IV presents a sample comparison of text with and without stopwords.

After stopword removal, separate word clouds are generated for clickbait and non-clickbait labels. The clickbait word cloud highlights frequently used sensational words designed to attract readers, while the non-clickbait word cloud displays more informative and neutral terms. This visualization clearly illustrates the linguistic differences between the two categories, helping to understand their distinct characteristics as shown in Fig. 2.

Label distribution analysis indicates a moderate imbalance in CLICK-ID, with 5,297 non-clickbait and 3,316 clickbait headlines. This imbalance can impact prediction results, especially in detecting the less frequent clickbait headlines, which motivates the use of data balancing techniques during data preparation. Maximum token length was also profiled to guide the choice of sequence length. As shown in Fig. 3, non-clickbait titles peak near 60 tokens, whereas clickbait titles exhibit a broader spread with a tail approaching 120 tokens. This distribution increases truncation risk and can complicate classification.

C. Balancing Data

The label distribution for each dataset variants is shown in Table VI. The original unbalanced dataset contains 5,297 non-clickbait and 3,316 clickbait headlines. The augmented dataset increases the clickbait samples to 5,298 while maintaining the original non-clickbait count, resulting in a balanced distribution of 10,595 samples. The oversampling technique duplicates minority class samples to match the majority, producing an equal of 5,297 headlines in each class with a total of 10,594 samples. In contrast, undersampling reduces the majority class to 3,316 headlines to match the minority class, yielding a balanced dataset of 6,632 samples. These balanced variants reduce class bias and are expected to improve clickbait headlines detection.

TABLE II
MODEL VARIATION

Model	Classifier Model	Dataset Type
Model #1	BERT	Unbalanced
Model #2	BERT	Augmented
Model #3	BERT	Oversampled
Model #4	BERT	Undersampled
Model #5	DistilBERT	Unbalanced
Model #6	DistilBERT	Augmented
Model #7	DistilBERT	Undersampled
Model #8	DistilBERT	Oversampled

TABLE III
DATASET AFTER DATA PREPROCESSING

No.	Title	Label	Label_score
1	*bi kenalkan standarisasi qr di kalimantan selatan*	Non-clickbait	0
2	*disindir hilang setelah masuk istana ini komentar teten masduki hingga johan budi*	Clickbait	1
3	*cepat kirim surat presiden revisi uu kpk ke dpr ini alasan jokowi*	Clickbait	1
4	*ribuan smp di jateng jadi pilot project gerakan menabung nasional ojk*	Non-clickbait	0

TABLE IV
COMPARISON OF DATA WITH AND WITHOUT STOPWORDS

No.	With Stopword	Without Stopword
1	*jenguk bj habibie di rspad kepala bppt beliau semangat recovery*	*jenguk bj habibie rspad kepala bppt beliau semangat recovery*
2	*sebuah mobil tertimpa pohon di pondok indah pengemudi terluka*	*mobil tertimpa pohon pondok indah pengemudi terluka*
3	*demokrat akan bangun museum dan art gallery sbyani di pacitan*	*demokrat bangun museum art gallery sbyani pacitan*

Fig. 2 Wordcloud for each label

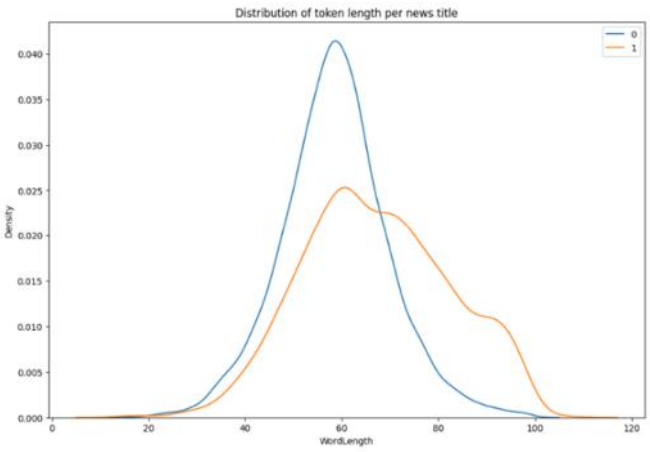

Fig. 3 Visualization of token length distribution per news headline

TABLE VI
LABEL DISTRIBUTION ACROSS DIFFERENT DATASET VARIATIONS

Dataset Type	Label Data		Total
	Non-Clickbait (0)	Clickbait (1)	
Unbalanced	5,297	3,316	8,613
Augmented	5,297	5,298	10,595
Oversampled	5,297	5,297	10,594
Undersampled	3,316	3,316	6,632

D. Splitting Data

The dataset variations were then split using an 80:10:10 ratio, ensuring a balanced approach to training, validation, and testing. 80% of the data is allocated for training, allowing the model to learn strong patterns effectively. 10% is used for validation to fine-tune hyperparameters and prevent overfitting, while the remaining 10% is for testing to evaluate the model's performance. Table VII shows the data distribution across different dataset variations used in this study.

E. Implementation of BERT model and DistilBERT model

This study fine-tunes two Indonesian transformer from the Hugging Face repository, bert-base-indonesian-1.5G and distilbert-base-indonesian. The models were pre-trained on about 522 MB of Indonesian Wikipedia, lowercased and tokenized with WordPiece using a 32,000-token vocabulary. Headlines are tokenized with the corresponding tokenizer and fed with a maximum sequence length of 120, batch size 32, and learning rate 1e-6. Training runs for up to 100 epochs with early stopping and a patience value of 2 to limit overfitting. Table VIII shows the parameter values used in this study.

After data formatting, BERT and DistilBERT are fine-tuned to classify clickbait and non-clickbait. Training runs for multiple epochs with the Adam optimizer and a fixed learning rate. Loss is computed using Sparse Categorical Crossentropy from logits, and accuracy is measured with Sparse Categorical Accuracy for the binary labels. A validation set monitors performance and early stopping halts training when improvement stalls.Table IX presents the training results for different classifier models and dataset types.

The BERT models trained on unbalanced and augmented datasets completed training in 7 epochs, while the oversampled and undersampled versions required 11 and 8 epochs, respectively. In contrast, DistilBERT models required more epochs to achieve similar or better performance, with the unbalanced dataset taking 10 epochs, and the augmented, oversampled, and undersampled datasets requiring 17, 19, and 12 epochs, respectively. This is expected, as DistilBERT has fewer layers than BERT, requiring more training iterations to reach comparable performance.

Table X presents a comparison of BERT and DistilBERT performance across different dataset balancing techniques. The results show that oversampling yields the highest evaluation scores for both models. BERT trained on the oversampled dataset achieves 93% in accuracy, precision, recall, and F1-score, outperforming other variations. DistilBERT with oversampling performs even better, reaching 94% across all metrics, demonstrating its effectiveness in improving classification performance.

TABLE VII
SPLITTING DATA RATIO

Dataset Type	Splitting Data		
	Train (80%)	Test (10%)	Val (10%)
Unbalaned	6,890	861	862
Augmented	8,476	1,059	1,060
Oversampled	8,475	1,059	1,060
Undersampled	5,305	663	664

TABLE VIII
HYPERPARAMETER

Parameter	Value
Max sequence length	120
Batch size	32
Learning rate	1e-6
Max epoch	100
Early stopper patience	2

TABLE IX
TRAINING MODEL RESULT

No. Model	Classifier Model	Dataset Type	Total Epoch
Model #1	BERT	Unbalanced	7 epoch
Model #2	BERT	Augmented	7 epoch
Model #3	BERT	Oversampled	11 epoch
Model #4	BERT	Undersampled	8 epoch
Model #5	DistilBERT	Unbalanced	10 epoch
Model #6	DistilBERT	Augmented	17 epoch
Model #7	DistilBERT	Oversampled	19 epoch
Model #8	DistilBERT	Undersampled	12 epoch

TABLE X
MODEL RESULT COMPARISON

Model	Dataset Type	Precision	Recall	F1-Score	Accuracy
BERT	Unbalanced	91%	90%	90%	91%
BERT	Augmented	92%	92%	92%	92%
BERT	**Oversampled**	**93%**	**93%**	**93%**	**93%**
BERT	Undersampled	92%	92%	92%	92%
DistilBERT	Unbalanced	91%	90%	90%	91%
DistilBERT	Augmented	93%	93%	93%	93%
DistilBERT	**Oversampled**	**94%**	**94%**	**94%**	**94%**
DistilBERT	Undersampled	91%	91%	91%	91%

Meanwhile, models trained on augmented and undersampled data achieve slightly lower but solid results, with accuracy and F1 around 92–93%. Unbalanced training yields the weakest scores, especially recall and F1, confirming the negative impact of class imbalance. Despite a potential overfitting risk, oversampling delivers the highest performance. BERT and DistilBERT trained on oversampled data provide the strongest accuracy and F1, making this configuration the preferred choice for CLICK-ID dataset.

To further analyze model, Fig. 4 presents the confusion matrix for BERT on the oversampled dataset. Of 1,060 predictions, it correctly classified 518 of 547 non-clickbait and 470 of 513 clickbait. Misclassifications include 29 false positives and 43 false negatives, slightly more missed clickbait when cues are subtle. By contrast, headlines with loud markers such as "*viral*", "*heboh*", "*bikin kaget*", or "*ternyata*" are typically flagged correctly.

As the best-performing model, the confusion matrix for DistilBERT with oversampled dataset is presented in Fig. 5. Of 1,060 total predictions, the model correctly classified 515 of 547 non-clickbait samples and 482 of 513 clickbait samples. Misclassifications are balanced, with 32 non-clickbait samples incorrectly labeled as clickbait (false positives) and 31 clickbait samples mislabeled as non-clickbait (false negatives). Compared with BERT, DistilBERT reduces missed clickbait while keeping false positives similar, indicating better sensitivity to context. A small set of borderline titles with ambiguous or sensational wording remains difficult. Overall, the matrices suggest that oversampling benefits both models, with DistilBERT delivering the most balanced performance across classes.

Table XI compares the proposed models with previous studies on clickbait detection. In [9], CNN and Bi-LSTM trained on unbalanced data achieved 88% and 87% accuracy, respectively. A multilingual BERT configuration with undersampling reached 91% [11], while a Bi-LSTM with Word2Vec and random oversampling achieved 89% [12]. In contrast, our proposed models deliver higher scores, with BERT reached 91% on the unbalanced dataset and 93% with oversampling, while DistilBERT matched 91% on the unbalanced dataset and attained 94% with oversampling, the best overall. These results indicate that class balancing improves performance and that transformer-based architectures, particularly DistilBERT, provide the strongest gains on the CLICK-ID dataset.

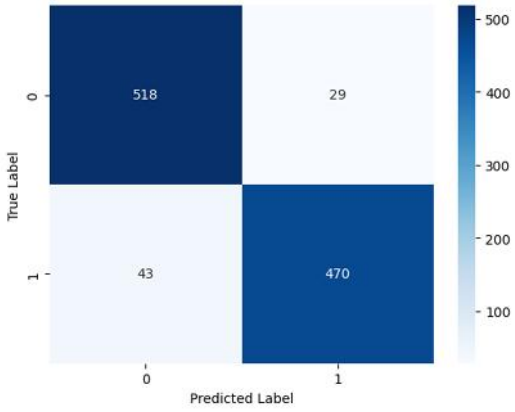

Fig. 4 Confusion matrix of BERT model with oversampled dataset

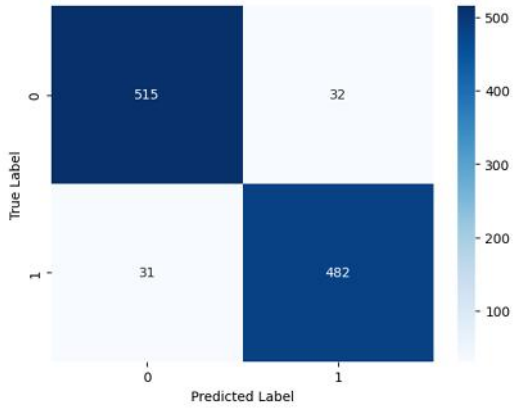

Fig. 5 Confusion matrix of DistilBERT model with oversampled dataset

TABLE XI
COMPARISON WITH PREVIOUS STUDY

Reference	Model	Dataset Type	Accuracy
[9]	CNN	Unbalanced	88%
	Bi-LSTM	Unbalanced	87%
[11]	Multilingual BERT	Undersampled	91%
[12]	Bi-LSTM+Word2Vec	Oversampled	89%
Proposed Model	BERT	Unbalanced	91%
	BERT	Oversampled	93%
	DistilBERT	Unbalanced	91%
	DistilBERT	**Oversampled**	**94%**

The success of BERT and DistilBERT in this study is attributed to their ability to process data in parallel and analyze text bidirectionally, allowing them to understand word context from both sides of a sentence through attention mechanisms. Unlike CNN, which processes data unidirectionally, and Bi-LSTM, which relies on sequential information, BERT captures deeper word relationships, leading to more accurate classification. This advantage makes BERT highly efficient for NLP tasks, particularly in complex text classification such as clickbait detection.

In Indonesian NLP, rich morphology and variable headline forms, including affixation, reduplication, clitics, compounding, flexible word order, and informal spelling, can hinder models that rely on fixed surface forms [27]. Transformer-based models such as BERT and DistilBERT mitigate this through subword tokenization and bidirectional contextual encoding, which improve coverage of rare or morphologically rich variants and capture long-range dependencies [5, 28]. On Indonesian corpora, pretrained Indonesian BERT variants handle out-of-vocabulary and code-mixed tokens more robustly, consistent with the lower error rates observed for subtle clickbait cases in the experiments [29]. These characteristics help the models generalize across affixed variants and compact headline styles and they support a stronger balance between false positives and false negatives in clickbait detection.

IV. CONCLUSION

Based on the conducted research, BERT trained on the oversampled dataset, achieved 93% for precision, recall, F1-score, and accuracy, while DistilBERT with oversampled dataset reached 93%. The oversampled data were produced by duplicating minority-class samples to balance the label distribution. These outcomes indicate that transformer-based models outperform RNN baselines such as CNN and Bi-LSTM. A key limitation is that the transformer models were tuned manually, which may be suboptimal. Future work could apply systematic hyperparameter tuning and evaluate additional transformer variants such as RoBERTa, which strengthens BERT pre-training through dynamic masking and extended training to further improve accuracy and robustness [30].

ACKNOWLEDGEMENT

The researcher sincerely appreciates the support and contributions of all the individuals and institutions involved in this study. The guidance, constructive feedback and valuable insights provided have been instrumental in shaping this research. Deep gratitude is expressed for the encouragement and participation throughout the process.

REFERENCES

[1] A. Agrawal, “Clickbait detection using deep learning,” in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 268–272. doi: 10.1109/NGCT.2016.7877426.

[2] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, “Stop Clickbait: Detecting and preventing clickbaits in online news media,” in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 2016, pp. 9–16. doi: 10.1109/ASONAM.2016.7752207.

[3] K. Scott, “You won’t believe what’s in this paper! Clickbait, relevance and the curiosity gap,” *J. Pragmat.*, vol. 175, pp. 53–66, Apr. 2021, doi: 10.1016/J.PRAGMA.2020.12.023.

[4] V. Kumar, D. Khattar, S. Gairola, Y. Kumar Lal, and V. Varma, “Identifying Clickbait: A multi-strategy approach using neural networks,” in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, in SIGIR ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1225–1228. doi: 10.1145/3209978.3210144.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers

for Language Understanding," 2019. doi: 10.48550/arXiv.1810.04805.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013, doi: 10.1109/TPAMI.2013.50.

[8] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, in SIGIR '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 115–124. doi: 10.1145/3077136.3080834.

[9] A. William and Y. Sari, "CLICK-ID: A novel dataset for Indonesian clickbait headlines," *Data Br.*, vol. 32, p. 106231, Oct. 2020, doi: 10.1016/J.DIB.2020.106231.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 5999–6009, 2017.

[11] M. N. Fakhruzzaman, S. Z. Jannah, R. A. Ningrum, and I. Fahmiyah, "Flagging clickbait in Indonesian online news websites using fine-tuned transformers," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 3, pp. 2921–2930, 2023.

[12] P. R. Togatorop, A. M. F. Tarigan, A. H. P. Sinaga, and E. P. D. Sidabutar, "Using deep learning and word embedding to detect clickbait in Indonesian headline news," in *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*, 2023, pp. 1–6. doi: 10.1109/ICoSNIKOM60230.2023.10364558.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, and T. Le Scao, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," 2020, [Online]. Available: https://arxiv.org/abs/1910.03771

[14] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3645–3650. doi: 10.18653/v1/P19-1355.

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[16] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, and M. Blokland, "The effects of data balancing approaches: A case study," *Appl. Soft Comput.*, vol. 132, p. 109853, Jan. 2023, doi: 10.1016/J.ASOC.2022.109853.

[17] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 6382–6388. doi: 10.18653/v1/d19-1670.

[18] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using Feature Selection," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1388–1400, 2010, doi: 10.1109/TKDE.2009.187.

[19] G. O. Assunção, R. Izbicki, and M. O. Prates, "Is augmentation effective to improve prediction in imbalanced text datasets?," 2023. [Online]. Available: http://arxiv.org/abs/2304.10283

[20] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: https://doi.org/10.1016/j.neunet.2018.07.011.

[21] A. J. Keya, M. A. H. Wadud, M. F. Mridha, M. Alatiyyah, and M. A. Hamid, "AugFake-BERT: Handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification," *Appl. Sci.*, vol. 12, no. 17, 2022, doi: 10.3390/app12178398.

[22] G. Liu, Y. Yang, and B. Li, "Fuzzy rule-based oversampling technique for imbalanced and incomplete data learning," *Knowledge-Based Syst.*, vol. 158, pp. 154–174, Oct. 2018, doi: 10.1016/J.KNOSYS.2018.05.044.

[23] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowledge-Based Syst.*, vol. 248, 2022, doi: 10.1016/j.knosys.2022.108839.

[24] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *J. Big Data*, vol. 8, no. 1, p. 101, 2021, doi: 10.1186/s40537-021-00492-0.

[25] N. Rai, D. Kumar, N. Kaushik, C. Raj, and A. Ali, "Fake news classification using transformer based enhanced LSTM and BERT," *Int. J. Cogn. Comput. Eng.*, vol. 3, pp. 98–105, Jun. 2022, doi: 10.1016/j.ijcce.2022.03.003.

[26] C. Wirawan, "Indonesian BERT base model (uncased)," Hugging Face. Accessed: Jul. 14, 2023. [Online]. Available: https://huggingface.co/cahya/bert-base-indonesian-1.5G

[27] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung, T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," 2023. [Online]. Available: https://arxiv.org/abs/2205.15960

[28] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and

pre-trained language model for Indonesian NLP," *arXiv Prepr. arXiv2011.00677*, 2020.

[29] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," 2021. [Online]. Available: https://arxiv.org/abs/2109.04607

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019. [Online]. Available: http://arxiv.org/abs/1907.11692