# Anda Tidak akan Percaya Apa yang Dapat Diselesaikan oleh *Clustering* dan *Data Augmentation* dalam Pembuatan Model *Clickbait Classifier*

**Darrel Danadyaksa Poli[1], Edbert Halim[2], Patrick Samuel Evans Simanjuntak[3], Andi Pujo Rahadi[4]**

[1]Ilmu Komputer Universitas Indonesia, Depok, 16424, email: darrel.danadyaksa@ui.ac.id
[2]Ilmu Komputer Universitas Indonesia, Depok, 16424, email: edbert.halim@ui.ac.id
[3]Ilmu Komputer Universitas Indonesia, Depok, 14320, email: patrick.samuel@ui.ac.id

***Abstract*——*Clickbait* adalah artikel daring dengan judul menyesatkan yang sengaja dibuat untuk menarik pembaca untuk membuka halaman dari berita tersebut. Adanya unsur *clickbait* pada suatu judul berita dapat menyebabkan disinformasi pada masyarakat dengan minat membaca yang rendah. Penelitian ini bertujuan untuk membuat sebuah model untuk menentukan apakah suatu judul berita mengandung unsur *clickbait* atau tidak serta menguji model apa yang paling akurat untuk mendeteksi unsur *clickbait* yang ada pada suatu judul berita. Penelitian ini mengandung pengujian akurasi penentuan apakah suatu judul berita terindikasi mengandung *clickbait* atau tidak dengan menggunakan empat model , yaitu Indobert, XGBoost, Catboost, serta Naive-Bayes. Model paling akurat yang kami buat dan temukan adalah Indobert dengan *accuracy score* = 0.83 dan *recall score* = 0.3 setelah model IndoBERT melalui proses *fine-tuning*. Selain menggunakan IndoBERT, kami juga melkaukan *clustering* dan *data augmentation* pada penelitian ini untuk mencoba kemungkinan lain dalam membantu *labeling* data dan memperbanyak data yang dapat digunakan.**

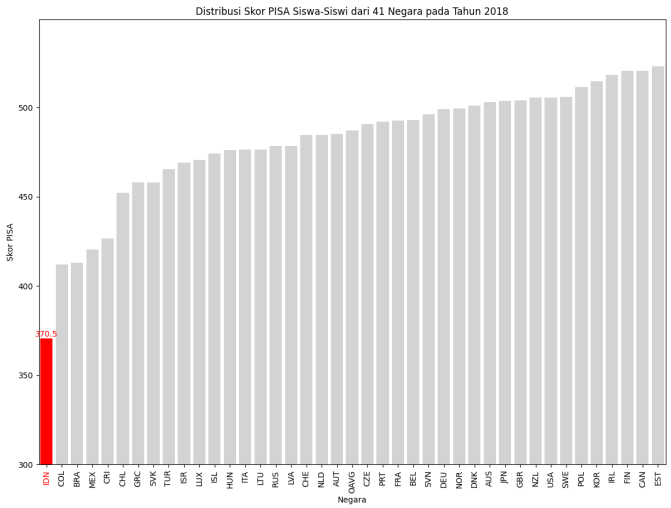


Fig. 1. Perbandingan Skor Pisa Indonesia dengan Negara Lain

KATA KUNCI

*Clickbait*, IndoBERT, XGBoost, Catboost, Naive-Bayes, *Clustering*, *Data Augmentation* Natural Language Processing

## I. PENDAHULUAN

*1) Latar Belakang:* Ketersediaan informasi di Indonesia melalui sosial media berkembang dengan pesat belakangan ini. Bahkan berdasarkan survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pengguna internet di Indonesia telah mencapai 215,64 juta orang pada periode 2022-2023. Tentunya dengan akses internet yang semakin besar, akses masyarakat Indonesia terhadap informasi semakin meningkat. Akan tetapi, akses terhadap informasi yang ada tidak berbanding lurus dengan tingkat literasi yang ada di Indonesia. Berdasarkan data tes PISA tahun 2018 yang dilansir dari dilakukan oleh Organisation for Economic Co-operation and Development (OECD), Indonesia menempati peringkat terakhir dari 41 negara dalam kemampuan membaca.

Hasil ini tentu saja memperlihatkan bahwa tingkat literasi Indonesia masih kurang baik. Ketersediaan internet membuat masyarakat Indonesia mulai meninggalkan media cetak dan beralih ke media digital. Banyaknya pembaca berita pada media digital membuat persaingan antar penulis berita digital. Untuk mendapatkan atensi dan pembaca, penulis berita *online* seringkali membuat suatu judul berita *online* yang mengandung unsur *clickbait*. *Clickbait* adalah suatu bentuk konten palsu yang sengaja dibuat untuk menarik perhatian calon pembaca dan membuat mereka penasaran agar mereka membaca konten yang ada [1]. Judul *clickbait* adalah judul yang menarik calon pembaca untuk mengklik suatu tautan dengan membangkitkan rasa penasaran calon pembaca [2]. Judul berita yang mengandung unsur *clickbait* biasanya tidak menampilkan isi berita seutuhnya. Hal tersebut dapat menjadi hal yang berbahaya jika para pembaca berita memiliki tingkat literasi yang kurang, terutama bagi masyarakat di Indonesia yang memiliki tingkat literasi yang sangat rendah. Kurangnya tingkat literasi menyebabkan seseorang hanya membaca judul berita saja dan menganggapnya sebagai kebenaran mutlak tanpa membaca keseluruhan dari beritanya terlebih dahulu. Hal ini dapat menyebabkan disinformasi secara massal yang berpotensi menimbulkan berbagai isu sosial yang merugikan [3]. Keresahan inilah yang membuat kami ingin mewujudkan sebuah *clickbait classifier* untuk ada di tengah-tengah sosial media untuk mencegah hal tersebut.

Berikut adalah contoh dari judul berita *online* yang mengandung unsur *clickbait* dan tidak mengandung unsur *clickbait*: Contoh judul berita *clickbait*: Alesandro Del Piero Ditangkap usai Tepergok Jual Narkoba di Mu-

ratara, Sabu 101 Gram Disita. Contoh judul berita *non-clickbait*: Sandiaga Salam Komando dengan Airlangga saat Nonton RI Vs Argentina. Judul berita pertama terlihat jelas mengundang rasa penasaran calon pembaca dengan membuat seolah-olah orang yang tertangkap adalah legenda sepakbola Italia, padahal orang tersebut hanyalah Warga Negara Indonesia (WNI) yang memiliki nama serupa dengan legenda sepakbola Italia, Alessandro Del Piero. Judul berita kedua menunjukkan hal yang sebenarnya terjadi yaitu dengan menjelaskan bahwa Sandiaga melakukan salam komando dengan Airlangga tanpa berusaha membangkitkan rasa penasaran calon pembaca. *Clickbait* perlu dicegah karena hal tersebut memanfaatkan *information gap* pada judul untuk memicu rasa penasaran pada para pembaca yang seringkali justru menyesatkan para pembaca [4]. Perhatikan bahwa *clickbait* memanfaatkan *cognitive bias* dalam otak manusia yang dapat dipengaruhi dengan mudah oleh media [5].

*2) Tujuan Penelitian:* Tujuan dari penelitian mengenai *clickbait classifier* ini adalah untuk membuat beberapa model bahasa untuk mendeteksi unsur *clickbait* pada suatu judul berita serta menentukan model bahasa apa yang paling akurat untuk menentukan apakah suatu judul berita mengandung unsur *clickbait* atau tidak.

*3) Manfaat Penelitian:* Manfaat dari penelitian ini bagi peneliti adalah menambah pengetahuan dan pengalaman mengenai model-model di bidang *NLP (Natural Language Processing)*. Manfaat dari penelitian ini bagi masyarakat adalah membantu mengurangi disinformasi yang kerap kali terjadi di Indonesia, terutama pada masyarakat yang memiliki tingkat literasi yang rendah.

*4) Batasan Penelitian:* -Penelitian ini hanya meneliti dan menentukan suatu model yang paling akurat dalam menentukan apakah suatu judul berita dalam bahasa Indonesia mengandung unsur *clickbait* atau tidak.

## II. Kajian yang Relevan

Di Indonesia, ada beberapa penelitian yang telah mencoba untuk membuat sebuah model yang dapat melakukan tugas klasifikasi berita. Sebuah penelitian [6] mencoba untuk melakukan klasifikasi terhadap judul *clickbait* dengan melakukan *similiarity scoring* dan juga mempertimbangkan ringkasan dari konten berita. Pada penelitian tersebut digunakan model IndoBERT. Penelitian tersebut berbeda dengan penelitian yang telah dilakukan [7] yang memerhatikan frekuensi sebuah kata muncul, nilai posisi kalimat, kesamaan dengan judul, panjang kalimat, pengurangan kalimat, dan peringkat kalimat. Tidak hanya kedua penelitian tersebut, ada penelitian lain [8] yang menggunakan *Multilingual Bidirectional Encoder Representations from Transformers* (M-BERT). Penelitian yang telah dilakukan ini menggunakan data dari CLICK-ID [9] yang juga menjadi data pada penelitian lain yang telah dilakukan [6].Penelitian lain juga dilakukan oleh [10] yang menggunakan model texitMultinomial Naïve Bayes untuk melakukan klasifikasi judul berita. Selain itu, digunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk melakukan pembobotan kata.

Setelah mengetahui penelitian-penelitian yang pernah dilakukan untuk menyelesaikan masalah *clickbait* di Indonesia, perlu diketahui juga mengetahui beberapa hal mengenai *natural language processing* (NLP). *Attention is All You Need* [11] adalah sebuah sebuah penelitian yang sangat penting di dunia *natural language processing* (NLP). Penelitian ini merupakan sebuah batu loncatan pada bidang NLP karena merupakan sebuah solusi baru untuk permasalahan yang ditimbulkan oleh arsitektur *Recurrent Neural Network* (RNN) dan *Long short-term memory* (LSTM) untuk mengolah data berurutan. Selain itu, data juga bisa dimasukkan ke dalam arsitektur transformer secara paralel. Perlu dicatat bahwa RNN dapat menimbulkan masalah *exploding gradient* dan *vanishing gradient* [12], tetapi masalah tersebut diselesaikan dengan menggunakan LSTM. Hal lain yang diperbaiki pada LSTM adalah penangkapan fitur laten dan data yang lebih baik. Perhatikan bahwa BERT [13] menggunakan model *transformer*. Model *transformer* inilah yang didasari oleh *Attention is All You Need* [11].

## III. Solusi Usulan

### A. Dataset

Penelitian ini menggunakan data berupa berita-berita *online* di Indonesia, mulai dari Detik.com, Fimela, Kompas.com, Liputan6, Okezone, Posmetro Medan, Republika, Sindonews, Tempo, Tribunnews, dan Wowkeren yang dipublikasi selama empat tahun terakhir, yaitu dari tahun 2019 hingga tahun 2023. Data-data yang digunakan didapatkan dari laman Mendeley Data yang telah dilakukan *labeling* secara manual dengan hanya memperhatikan judul saja. Kita juga melakukan *scrapping* dengan menggunakan kode python untuk mengambil judul dan isi dari laman-laman berita *online* secara satu per satu. Alasan penggunaan data melalui teknik *scrapping* dan data dari laman Mendeley Data adalah karena data tersebut merupakan berita-berita aktual dan faktual yang ada di Indonesia, sehingga model dapat di-*train* dengan data yang sesuai dengan kondisi nyata yang ada di laman-laman berita *online*.

### B. Visualisasi Dataset

Berikut adalah data yang gunakan, dimana terdapat 14757 judul berita dalam data yang digunakan [9]. Perhatikan bahwa data ini merupakan data yang diberikan label oleh tiga peneliti berbeda. Hal ini menyebabkan reabilitas data yang tidak sempurna [8]. Walaupun begitu, penelitian ini tetap melanjutkan menggunakan data tersebut untuk melihat bagaimana model yang akan dibuat dapat menerima data yang tidak sempurna tersebut. Berikut adalah beberapa gambaran mengenai data yang digunakan dalam penelitian ini. Gambar pertama menunjukan jumlah judul berita yang mengandung unsur *clickbait* dan tidak mengandung unsur *clickbait* berdasarkan portal berita daring.

Setelah melihat mengenai distribusi data yang *clickbait* dan *non-clickbait* berdasarkan judul, kita juga dapat melihat perbandingan data *clickbait* dan *non-clickbait* berdasarkan penerbit berita. Disini, perbandingan tersebut dibandingkan dengan banyaknya data berita yang

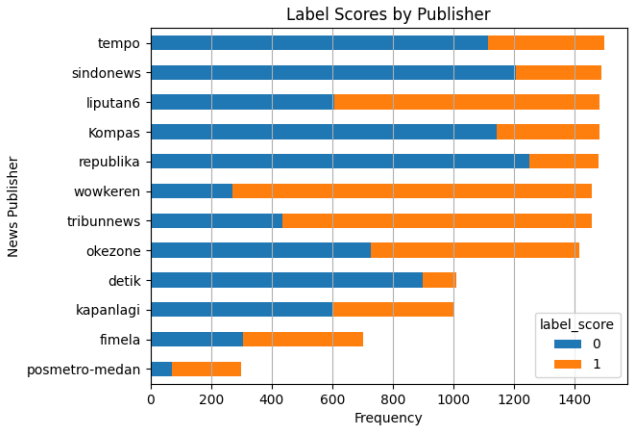


Fig. 2. Jumlah Data *Clickbait/Non-Clickbait pada Masing-masing Portal Berita Daring*

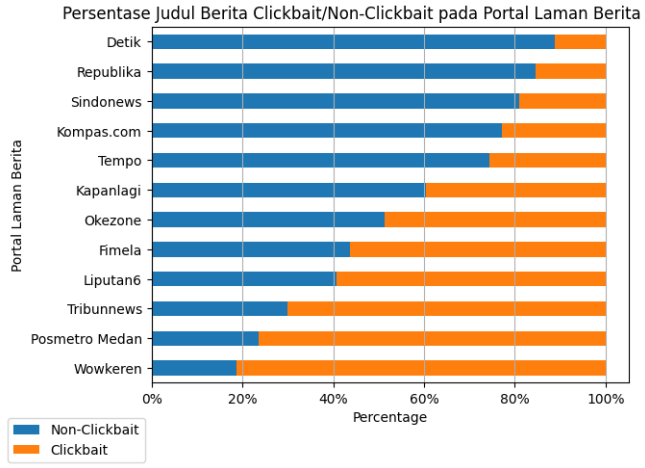


Fig. 3. Persentase Berita Perbandingan *clickbait/non-clickbait* pada masing-masing portal berita

diterbitkan oleh setiap penerbit. Dengan begitu, diharapkan data yang dihasilkan dapat menjadi lebih rata untuk dilihat.

Perhatikan juga hal lain menarik dari data yang digunakan dalam penelitian ini, yaitu persebaran data dari *feature engineering* yang dilakukan. Jika kita menganalisis judul berita mana yang *clickbait*, kita akan mendapati bahwa judul berita dengan tanda tanya dan tanda seru pasti merupakan *clickbait*.

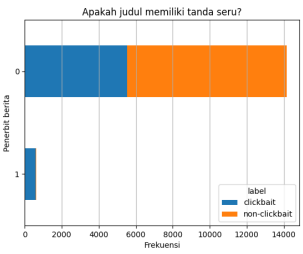


Fig. 4. Persebaran berita yang *clickbait/non-clickbait* jika judul memiliki tanda tanya

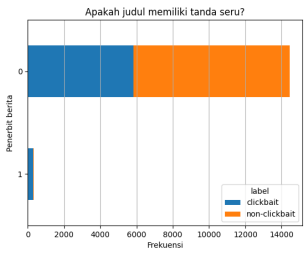


Fig. 5. Persebaran berita yang *clickbait/non-clickbait* jika judul memiliki tanda seru

### C. Metode yang Digunakan

*1) Data Pre-processing:* Pada tahap *Data Pre-processing*, data akan diberikan beberapa perlakuan sebelum dimasukkan ke dalam model *Machine Learning*. Perlakuan ini penting untuk dilakukan pada setiap model yang digunakan karena adanya *noise* pada data. Oleh karena itu, teknik ini digunakan untuk memperoleh jumlah data, struktur, dan format yang terbaik [14]. Teknik *pre-processing* yang digunakan untuk setiap model perlu dikhususkan. Hal ini dikarenakan dari perbedaan algoritma yang digunakan dalam setiap model yang digunakan. Walaupun begitu, *pre-processing* yang dilakukan oleh semua model adalah penghapusan kolom *label, url, date,* dan *time*. Selanjutnya, kolom *sub-category* pada data akan dikenakan perlakuan *data generalization*, yaitu dengan menggabungkan beberapa *sub-category* yang mirip, seperti *sub-category "startup"* dan *"technology"* digabung menjadi *sub-category "technology"*. Tahapan *data pre-processing* selanjutnya dilakukan secara terpisah sesuai dengan model masing-masing.

Teknik *pre-proccessing* yang diberikan kepada model XGBoost, Catboost, dan Naive-Bayes adalah dengan melakukan *text cleaning* dengan menghapus emoji, simbol, bendera, nomor, dan semua karakter pada judul yang bukan merupakan huruf. Selanjutnya teks pada kolom *title* akan dikenakan vektorisasi dengan menggunakan *CountVectorizer*. *CountVectorizer* adalah suatu alat yang digunakan untuk memecah suatu kalimat utuh menjadi kata-kata terpisah lalu menghitung frekuensi kata tersebut muncul pada suatu kalimat.

Teknik *pre-processing* yang digunakan untuk model IndoBERT *text cleaning, tokenization*, dan pembuatan *attention mask*. Pada *text cleaning*, huruf semua huruf pada judul dijadikan huruf kecil. Selain itu, angka dan *HTML Tags* dihapus dari judul. Tidak hanya itu, *Uniform Resource Locator* (URL), *emoticon*, simbol dan *picograph*, *transport* dan *map symbol*, dan *flags*, *emoji* dihapus dari judul-judul yang ada pada data

*2) Model:* Model pertama yang akan digunakan adalah model eXtreme Gradient Boosting (XGBoost). Model *XGBoost* dipilih karena kebutuhan teknik *parallel computing* [15] yang dibutuhkan dikarenakan banyaknya data yang dihasilkan dari *tokenization*.

Catboost digunakan dalam penelitian ini karena sifatnya yang memanfaatkan permutasi. Hal ini sangat dibutuhkan dalam model yang perlu dibuat karena permutasi merupakan salah satu inti dari *Clickbait*. Pada lain sisi, *Naïve Bayes* digunakan karena kita ingin peluang kondisional untuk bermain peran dalam model yang akan dibuat.

Model kedua terakhir yang digunakan dalam penelitian ini adalah IndoBERT. IndoBERT digunakan karena pertimbangan dari [6] yang dapat memeroleh hasil penelitian yang baik dengan menggunakan model tersebut. Model IndoBERT [16] diberikan *tokenization* dan *attention mask* agar dapat bekerja. Setelah itu, dilakukan *fine-tuning* pada IndoBERT untuk model *classifier clickbait* kami.

*3) Perbedaan dengan solusi sebelumnya:* Metode terakhir yang akan kami coba adalah *clustering*. Harapan kami adalah dengan metode ini, setiap *cluster* memiliki

sebuah label yang dominan, entah itu *clickbait* atau *non-clickbait*. Jika kami bisa menemukan model *clustering* yang memiliki label yang dominan di tiap *cluster*, kami dapat membuktikan bahwa *labeling* data dapat dilakukan dengan lebih mudah. Hal ini sangat krusial untuk dilakukan karena *labeling* data merupakan salah satu tahapan yang sangat menguras tenaga.

Selain itu, hal yang belum ditemukan dalam penelitian yang pernah dilakukan di Indonesia adalah *data augmentation*. *Data augmentation* adalah suatu teknik manipulasi data dengan menggunakan teknik artifisial untuk memperluas dataset dengan data yang ada. Beberapa contohnya dalam konteks Natural Language Processing (NLP) adalah dengan melakukan *back-translation* [17], melakukan penggantian kata dengan sinonimnya [18], melakukan penukaran kata secara acak [18], menghapus kata secara acak [18], mengubah susunan kalimat, serta menghapus kata duplikat.

*4) Evaluasi model:* Evaluasi model merupakan hal yang penting saat membuat sebuah model. Hal tersebut penting untuk dilakukan sebagai cara untuk melihat seberapa bagus model yang telah dibuat. Acuan yang sering digunakan antara lain adalah *accuracy, precision, recall*, serta *F1 score*. Kami menggunakan *accuracy score* sebagai salah satu acuan evaluasi karena *accuracy score* mencerminkan akurasi dari prediksi yang dilakukan secara umum, yaitu dengan membagi prediksi yang benar dengan keseluruhan data yang ada. Persamaan dari *accuracy score* adalah sebagai berikut:

$$\textit{Accuracy Score} = \frac{TP + TN}{TP + FP + FN + TN}$$

Keterangan:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Meskipun ada beberapa metrik yang dapat digunakan untuk menilai model secara kuantitatif, dalam penelitian ini diputuskan untuk menggunakan metrik *accuracy score* dan *recall*. Metrik *recall* digunakan untuk menilai model yang telah dibuat karena sifatnya yang memerhatikan *false negative* pada prediksi yang telah dibuat. *False negative* ini penting untuk diperhatikan untuk kasus spesifik yang sedang dibuat model *classifier*-nya, yaitu prediksi yang dibuat harus meminimalisir prediksi yang menandakan suatu judul berita sebagai *non-clickbait*, padahal berita tersebut merupakan berita *clickbait*. Hal ini perlu dihindari karena kelalaian dalam membuat prediksi berita *clickbait* menjadi *non-clickbait* artinya dapat meloloskan berita *clickbait*. Jika model yang telah dibuat ini di-*deploy*, maka pembaca berita dapat menemukan berita *clickbait* pada tempat pembaca berita tersebut membaca berita. Dengan pertimbangan tersebut, dapat dapat disimpulkan bahwa *recall* merupakan metrik yang perlu digunakan dalam penilaian model yang akan dibuat. Berikut adalah persamaan untuk *recall*:

$$\textit{Recall Score} = \frac{TP}{TP + FN}$$

Dimana:
TP = True Positive
FN = False Negative

## IV. Hasil Eksperimen dan Pengujian

Pengujian dilakukan dengan kurang lebih 4428 test data yang belum pernah dilihat oleh model sebelumnya. Hal ini dilakukan untuk melihat apakah model yang telah dibuat untuk menyelesaikan tugas klasifikasi ini bisa menyelesaikan tugas tersebut dengan baik. Terlihat terdapat perbedaan yang signifikan antara model IndoBERT yang telah dilakukan *fine-tuning* dengan model lainnya.

## V. Note

Berikut adalah lampiran kami: Link lampiran

TABLE I
Evaluasi Model

| Model | Accuracy | Recall |
|---|---|---|
| IndoBERT | 0.83 | 0.83 |
| XGBoost | 0.76 | 0.64 |
| CatBoost | 0.77 | 0.62 |
| Naive Bayes | 0.69 | 0.57 |

*Skor dapat meningkat sesuai dengan peningkatan *performance* model yang ada

## References


[1] D. Varshney and D. Vishwakarma, "A unified approach for detection of clickbait videos on youtube using cognitive evidences," *Applied Intelligence*, vol. 51, pp. 1–22, 07 2021.

[2] Şura Genç and E. Surer, "Clickbaittr: Dataset for clickbait detection from turkish news sites and social media with a comparative analysis via machine learning algorithms," *Journal of Information Science*, vol. 49, no. 2, pp. 480–499, 2023. [Online]. Available: https://doi.org/10.1177/01655515211007746

[3] K. Shu, S. Wang, D. Lee, and H. Liu, "Mining disinformation and fake news: Concepts, methods, and recent advancements," 2020.

[4] K. Scott, "You won't believe what's in this paper! clickbait, relevance and the curiosity gap," *Journal of Pragmatics*, vol. 175, pp. 53–66, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378216621000229

[5] S. Pandey and G. Kaur, "Curious to click it?-identifying clickbait using deep learning and evolutionary algorithm," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 1481–1487.

[6] H. Ahmadi and A. Chowanda, "Clickbait classification model on online news with semantic similarity calculation between news title and content," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, pp. 1986–1994, Mar. 2023. [Online]. Available: https://ejurnal.seminar-id.com/index.php/bits/article/view/3030

[7] B. W. Rauf, S. Raharjo, and H. Sismoro, "Deteksi clickbait dengan sentence scoring based on frequency," *Jurnal Teknologi Informasi*, vol. 4, no. 2, p. 247, December 2020. [Online]. Available: http://jurnal.una.ac.id/index.php/jurti/article/view/1381/1458

[8] M. N. Fakhruzzaman, S. Z. Jannah, R. A. Ningrum, and I. Fahmiyah, "Clickbait headline detection in indonesian news sites using multilingual bidirectional encoder representations from transformers (m-bert)," 2021.

[9] A. William and Y. Sari, "Click-id: A novel dataset for indonesian clickbait headlines," *Data in Brief*, vol. 32, p. 106231, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920311252

[10] F. A. Ramadhan, S. H. Sitorus, and T. Rismawan, "Penerapan metode multinomial naïve bayes untuk klasifikasi judul berita clickbait dengan term frequency - inverse document frequency," *Jurusan Informatika Universitas Tanjungpura*, vol. Vol 11, No 1 (2023), 2023. [Online]. Available: https://jurnal.untan.ac.id/index.php/justin/article/view/57452/75676596395

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
[12] R. C. Staudemeyer and E. R. Morris, “Understanding lstm – a tutorial into long short-term memory recurrent neural networks,” 2019.
[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.
[14] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, ser. Intelligent Systems Reference Library. Springer International Publishing, 2014. [Online]. Available: https://books.google.co.id/books?id=SbFkBAAAQBAJ
[15] T. Chen and C. Guestrin, “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. [Online]. Available: https://doi.org/10.11452F2939672.2939785
[16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp,” in *Proceedings of the 28th COLING*, 2020.
[17] S. Connor, T. M. Khoshgoftaar, and F. Borko, “Text data augmentation for deep learning,” *Journal of Big Data*, vol. 8, no. 1, 12 2021, copyright - © The Author(s) 2021. This work is published under http://creativecommons.org/licenses/by/4.0/ (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2023-03-09. [Online]. Available: https://www.proquest.com/scholarly-journals/text-data-augmentation-deep-learning/docview/2553125823/se-2
[18] J. Wei and K. Zou, “Eda: Easy data augmentation techniques for boosting performance on text classification tasks,” 2019.