KLASIFIKASI UJARAN KEBENCIAN *MULTI LABEL* MENGGUNAKAN ARSITEKTUR *LONG SHORT TERM MEMORY* DAN *TRANSFORMER* DENGAN *BACK TRANSLATION* DAN BERT

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar

Sarjana Matematika

Oleh:

PUTRI PRATIWI

NIM. 08011282126028

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SRIWIJAYA

2025

LEMBAR PENGESAHAN

KLASIFIKASI UJARAN KEBENCIAN *MULTI LABEL* MENGGUNAKAN ARSITEKTUR *LONG SHORT TERM MEMORY* DAN *TRANSFORMER* DENGAN AUGMENTASI *BACK TRANSLATION* DAN BERT

SKRIPSI

Sebagai Salah Satu Syarat untuk Memperoleh Gelar

Sarjana Matematika

Oleh

PUTRI PRATIWI

NIM. 08011282126028

Indralaya, 20 Maret 2025

Pembimbing Kedua	**Pembimbing Utama**
Dr. Bambang Suprihatin, S.Si., M.Si	**Dr. Anita Desiani, S.Si., M.Kom.**
NIP. 197101261994121001	**NIP.197712112003122002**

Mengetahui,
Ketua Jurusan Matematika

Dr. Dian Cahyawati Sukanda, S.Si., M.Si.
NIP. 197303212000122001

PERNYATAAN KEASLIAN KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama Mahasiswa	:	Putri Pratiwi
NIM	:	08011282126028
Jurusan	:	Matematika
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

Menyatakan bahwa skripsi ini adalah hasil karya ilmiah saya sendiri dan karya ilmiah ini belum pernah diajukan sebagai pemenuhan persaratan untuk memperoleh gelar kesarjanaan strata satu (S1) dari Universitas Sriwijaya maupun perguruan tinggi lain. Semua informasi yang dimuat didalam skripsi ini yang berasal dari penulis lain baik yang dipublikasi atau tidak telah diberikan penghargaan dengan mengutip nama sumber penulis baik yang secara benar. Semua isi dari skripsi ini sepenuhnya menjadi tanggung jawab saya sebagai penulis.

Demikianlah surat pernyataan ini saya buat dengan sebenarnya.

Indralaya, 20 Maret 2025

Penulis

Putri Pratiwi

NIM. 08011282126028

HALAMAN PERSEMBAHAN

Kupersembahkan skripsi ini untuk:

Yang Maha Kuasa Allah Subhanahu Wa Ta'ala

Kedua orang tuaku tercinta,

Adik-adikku tersayang,

Keluarga besarlu,

Semua guru dan dosenku,

Sahabat-sahabatku,

Almamaterku.

Diriku sendiri

Motto

"Tidak mudah bukan berarti tidak mungkin, kamu bisa lebih dari apa yang kamu kira"

-Putri Pratiwi

KATA PENGANTAR

Puji syukur atas kehadirat Allah Subhanahu Wa Ta'ala yang telah memberikan rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi berjudul "Klasifikasi Ujaran Kebencian *Multi Label* Menggunakan Arsitektur *Long Short Term Memory* dan *Transformer* dengan Augmentasi *Back Translation* dan BERT". Skripsi ini ditulis sebagai salah satu syarat memperoleh gelar sarjana sains studi Matematika di Fakultas MIPA Universitas Sriwijaya.

Penulis menyadari bahwa skripsi ini tidak mungkin terselesaikan tanpa adanya semangat, dukungan, bantuan, bimbingan dan nasihat yang diberikan berbagai pihak selama proses penyusunan ini berlangsung. Penulis mengucapkan terima kasih yang sebesar-besarnya kepada orang tua tercinta, Ayahku **Supriadi dan** Ibuku **Sri Wahyuningsih** yang tidak pernah berhenti berjuang dan memberikan yang terbaik untukku sebagai putrinya. Terima kasih karena tak pernah lelah mendidik, menasehati, membimbing, mendukung dan terus mendo'akan. Penulis juga ingin menyampaikan ucapan terima kasih kepada sebesar-besarnya dan penghargaan setinggi-tingginya kepada:

1. Bapak **Prof. Hermansyah, S.Si., M.Si., Ph.D** selaku Dekan Fakultas MIPA Universitas Sriwijaya.
2. Ibu **Dr. Dian Cahyawati Sukanda, S.Si., M.Si**. selaku Ketua Jurusan Matematika dan Ibu **Des Alwine Zayanti, S.Si., M.Si.** selaku Sekretaris Jurusan Matematika yang telah membimbing dan mengarahkan dalam urusan akademik selama menempuh perkuliahan di Jurusan Matematika FMIPA Universitas Sriwijaya.

3. Ibu **Dr. Anita Desiani, S.Si., M.Kom.** selaku Dosen Pembimbing Pertama dan Bapak **Dr. Bambang Suprihatin, S.Si., M.Si.** selaku Dosen Pembimbing Kedua yang telah bersedia meluangkan waktu, tenaga, dan pikiran untuk memberikan bimbingan, arahan dan didikan yang berharga selama pembuatan skripsi, perlombaan dan proses perkuliahan.
4. Bapak **Drs. Endro Setyo Cahyono, M.Si** dan Ibu **Irmeilyana, S.Si., M.Si** selaku dosen pembahas, telah memberikan respons, kritik, dan saran yang sangat berguna untuk perbaikan dan penyelesaian skripsi ini.
5. **Seluruh Dosen di Jurusan Matematika FMIPA** yang telah memberikan ilmu, nasihat, motivasi, serta bimbingan selama proses perkuliahan. Bapak **Irwansyah** dan Ibu **Hamidah** selaku staf administrasi di Jurusan Matematika FMIPA Universitas Sriwijaya yang telah membantu penulis selama perkuliahan.
6. **Himpunan Mahasiswa Matematika (Himastik)** yang telah menjadi wadah untuk mengembangkan minat, bakat, dan kreativitas di bidang matematika.
7. Semua sahabat seperjuangan **Komputasi 2021** selama masa perkuliahan dan proses skripsi. Kakak-kakak tingkat bidang komputasi yang telah membantu dan membagikan ilmunya kepada penulis, serta adik-adik tingkat yang telah membantu proses skripsi.
8. Saudaraku, keluarga besarku, serta sahabat-sahabatku yang senantiasa memberikan semangat dan doa terbaik untuk penulis.
9. Semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah memberikan bantuan dalam menyelesaikan skripsi ini hanya ucapan terima

kasih yang dapat penulis berikan.

Semoga skripsi ini dapat menambah pengetahuan dan bermanfaat bagi mahasiswa/i Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya dan semua pihak yang membutuhkan.

Indralaya, Maret 2025

Penulis

MULTI-LABEL HATE SPEECH CLASSIFICATION USING LONG SHORT TERM MEMORY AND TRANSFORMER ARCHITECTURE WITH BACK TRANSLATION AUGMENTATION AND BERT

***By*:**

PUTRI PRATIWI
NIM.08011282126028

ABSTRACT

Multi-label hate speech can be categorized by more than one label at once, such as individual, religious, racial hate speech, with different levels of severity. In the multi-label context, class and label are used to determine the categories in the data. Labels refer to the types of hate speech found in the data, while class refers to the combination of the various labels. One of the labeling methods that can be used to determine the labels in the data is label powerset. Multi-label hate speech spreads quickly and widely, so automatic early detection is needed. This research uses a combination of classification architecture and augmentation techniques. The classification architecture used is Long Short Term Memory (LSTM) to process the sequence of important information and Transformer to understand the context globally. The combination of architectures requires a large amount of data. The method used to multiply the data is to perform augmentation using the back translation method and Bidirectional Encoder Representation from Transformer (BERT). The results show an increase in the amount of data up to three times. The combination of architectures has good model performance. The accuracy result of 86% shows that the model is able to predict the class correctly as a whole. The precision result of 86% shows that the model can identify positive classes with a low error rate. The recall result of 85% shows that the model can detect most of the positive classes from the available data. The f1-score result of 85% shows that the model is consistent in classifying positive classes. In each class, the model performed very well in detecting the H0 and H26 classes with accuracy, precision, recall, and f1-score results of more than 90% each. In the H17, H23, and H31 classes, the model still struggled with classification. This research shows that the combination of LSTM and Transformer architecture as well as the combination of back translation augmentation and BERT can be used for multi-label hate speech classification. For further research, the balance of data between classes needs to be considered so that the model's performance is more optimal.

Keywords: hate speech, back translation, Bidirectional Encoder Representation from Transformer (BERT), Long Short Term Memory (LSTM), Transformer

KLASIFIKASI UJARAN KEBENCIAN *MULTI LABEL* MENGGUNAKAN ARSITEKTUR *LONG SHORT TERM MEMORY* DAN *TRANSFORMER* DENGAN AUGMENTASI *BACK TRANSLATION* DAN BERT

Oleh:

PUTRI PRATIWI
NIM.08011282126028

ABSTRAK

Ujaran kebencian *multi label* dapat dikategorikan lebih dari satu *label* sekaligus, seperti ujaran kebencian individu, agama, ras, dengan tingkat keparahan yang berbeda. Pada konteks *multi label,* kelas dan *label* digunakan untuk menentukan kategori pada data. *Label* merujuk pada jenis ujaran kebencian yang ditemukan dalam data, sementara kelas merujuk pada hasil kombinasi dari berbagai label tersebut. Salah satu metode pelabelan yang dapat digunakan untuk menentukan *label* pada data adalah *label powerset*. Ujaran kebencian *multi label* menyebar dengan cepat dan luas, sehingga diperlukan deteksi dini secara otomatis. Penelitian ini menggunakan kombinasi arsitektur klasifikasi dan teknik augmentasi. Arsitektur klasifikasi yang digunakan adalah *Long Short Term Memory* (LSTM) untuk memproses urutan informasi penting dan *Transformer* untuk memahami konteks secara global. Kombinasi arsitektur memerlukan jumlah data yang banyak. Cara yang digunakan untuk memperbanyak data adalah melakukan augmentasi menggunakan metode *back translation* dan *Bidirectional Encoder Representation from Transformer* (BERT). Hasil penelitian menunjukkan peningkatan jumlah data hingga tiga kali lipat. Kombinasi arsitektur memiliki kinerja model yang baik. Hasil akurasi sebesar 86% menunjukkan bahwa model mampu memprediksi kelas dengan benar secara keseluruhan. Hasil presisi sebesar 86% menunjukkan bahwa model dapat mengidentifikasi kelas positif dengan tingkat kesalahan yang rendah. Hasil *recall* sebesar 85% menunjukkan bahwa model dapat mendeteksi sebagian besar kelas positif dari data yang tersedia. Hasil *f1-score* sebesar 85% menunjukkan bahwa model konsisten dalam mengklasifikasikan kelas positif. Pada masing-masing kelas, model menunjukkan performa sangat baik dalam mendeteksi kelas H0 dan H26 dengan hasil akurasi, presisi, *recall,* dan *f1-score* masing lebih dari 90%. Pada kelas H17, H23, dan H31 model masih kesulitan dalam melakukan klasifikasi. Penelitian ini menunjukkan bahwa kombinasi arsitektur LSTM dan *Transformer* serta kombinasi augmentasi *back translation* dan BERT dapat digunakan untuk klasifikasi ujaran kebencian *multi label*. Untuk penelitian selanjutnya, keseimbangan data antar kelas perlu diperhatikan agar konerja model lebih optimal.

Kata kunci: ujaran kebencian, *back translation, Bidirectional Encoder Representation from Transformer* (BERT), *Long Short Term Memory* (LSTM), *Transformer*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Cuitan di media sosial seperti Twitter dapat berdampak negatif, salah satunya dalam bentuk ujaran kebencian. Ujaran kebencian memiliki berbagai tingkat keparahan, mulai dari rendah, sedang, hingga tinggi (Hana *et al*., 2020). Pada satu ujaran kebencian dapat masuk ke dalam lebih dari satu kategori, seperti ujaran kebencian individu, agama, atau ras, dengan tingkat keparahan yang berbeda sehingga perlu dilakukan pelabelan agar dapat memberi penangan sesuai kategorinya. Ujaran kebencian memiliki dampak terhadap negara berupa ketegangan sosial antar kelompok masyarakat (Elsafoury, 2023). Untuk mengatasi masalah tersebut, dilakukan klasifikasi otomatis menggunakan algoritma yang ada pada *deep learning*.

Transformer merupakan algoritma *deep learning* yang banyak digunakan untuk melakukan klasifikasi (Mozafari *et al.*, 2020). Salah satu keunggulan *Transformer* adalah mampu menangkap informasi global melalui mekanisme *self attention*, karena memproses *input* secara pararael (Kovaleva *et al*., 2019). Penerapan *Transformer* pada klasifikasi telah banyak dilakukan. Bilal *et al*. (2023) menggunakan arsitektur *Transformer* untuk klasifikasi ujaran kebencian bahasa Roman Urdu dengan akurasi, *f-measure*, presisi, dan *recall* sebesar 83%. Sarkar *et al*. (2021) menggunakan arsitektur *Transformer* untuk klasifikasi ujaran kebencian bahasa Inggris dengan *f1-score* sebesar 87%. *Transformer* adalah arsitektur yang

memiliki parameter besar sehingga membutuhkan jumlah data yang banyak. Jumlah data cuitan *tweet* masih terbatas terutama pada bahasa selain bahasa Inggris (Ibrohim *and* Budi., 2023). Jumlah data yang terbatas dengan parameter yang besar dapat menyebabkan pembelajaran *Transformer* menjadi *overfitting* (Tra *et al*., 2019).

Untuk mengatasi keterbatasan data terutama pada cuitan *tweet* adalah dengan melakukan teknik augmentasi (Shorten *et al*., 2021). Augmentasi merupakan teknik untuk memperbanyak data. Pada suatu teks, augmentasi dilakukan dengan cara menghapus, menyisipkan atau mengganti kata sehingga diperoleh kalimat baru yang lebih bervariasi (Sabty *et al*., 2021). Salah satu metode augmentasi pada teks adalah *Bidirectional Encoder Representations from Transformers* (BERT).

Augmentasi BERT adalah teknik untuk menambah jumlah data teks dengan menggunakan *self atttention* (Wu *et al*., 2019). *Self attention* pada BERT digunakan untuk menangkap hubungan antar kata dalam satu kalimat, sehingga dapat memperhatikan hubungan dua arah yaitu arah sebelum dan arah setelah kata yang di-*mask*. Augmentasi BERT tidak langsung menyisipkan kata baru, tetapi melakukan pembobotan menggunakan fungsi *softmax*. Kata dengan bobot tertinggi digunakan untuk mengganti kata yang di-*mask*. (Kolesnichenko *et al*., 2023). Kelebihan dari augmentasi BERT yaitu menghasilkan teks bervariasi dengan makna sesuai kalimat aslinya, karena penggantian kata dihasilkan dengan melakukan pembobotan nilai.

Beberapa penelitian telah melakukan augmentasi data menggunakan BERT. Kapil *and* Ekbal (2022) menerapkan augmentasi BERT pada *dataset* bahasa India menggunakan arsitektur *Transformer* dengan nilai *f1-score* sebesar 74%. Takawane *et al.* (2023) menerapkan augmentasi BERT pada *dataset* bahasa Inggris menggunakan arsitektur *Transformer* dengan nilai *f1-score* sebesar 72%. BERT banyak diterapkan dalam bahasa Inggris karena korpus pelatihannya berasal dari teks berbahasa Inggris sehingga, penerapan pada bahasa lain menjadi lebih rumit (Devlin *et al.*, 2019).

Penerapan augmentasi BERT pada bahasa selain bahasa Inggris, diperlukan proses penerjemahan menggunakan teknik *back translation*. *Back translation* merupakan teknik sederhana untuk menerjemahkan teks dari bahasa asal ke bahasa Inggris, kemudian menerjemahkannya kembali ke bahasa asal (Beddiar *et al.*, 2021). Desiani *et al.* (2023) melakukan augmentasi *back translation* dan EDA untuk klasifikasi ujaran kebencian pada dataset bahasa Indonesia menggunakan arsitektur klasifikasi *Transformer*. Hasil penelitian menunjukkan bahwa nilai akurasi, presisi, *recall*, dan *f1-score* sebesar 85% dan hanya menggunakan dua *label* yaitu *hate speech* dan *non hate speech*. Beddiar *et al.* (2021) melakukan augmentasi *back translation* dan *paraphrasing* untuk klasifikasi ujaran kebencian pada dataset bahasa Jerman dengan menggunakan arsitektur klasifikasi LSTM. Hasil penelitian menunjukkan bahwa nilai akurasi dan *f1-score* sebesar 99%, namun hanya melakukan klasifikasi pada dua kelas yaitu *hate* dan *non hate*. Penerapan metode augmentasi dapat memenuhi kebutuhan data *training* yang besar oleh arsitektur *Transformer* (Ansari *et al.*, 2021). Selain membutuhkan data yang besar,

Transformer tidak dapat menyeleksi informasi penting dan beresiko kehilangan penekanan informasi yang benar-benar penting (Vaswani *et al*., 2017).

Arsitektur yang dapat menyeleksi informasi penting adalah arsitektur *Long Short-Term Memory* (LSTM). Arsitektur LSTM mampu memilih informasi penting melalui mekanisme *gate* (Fazil *et al*., 2023). LSTM memiliki tiga *gate* utama yaitu *input gate, output gate*, dan *forgot gate*. *Input gate* berfungsi untuk menambahkan informasi baru yang akan disimpan pada tempat penyimpanan informasi. *Output gate* berfungsi untuk menghasilkan informasi dari pembelajaran model yang digunakan sebagai *output*. *Forgot gate* berfungsi untuk menyeleksi informasi penting (Ayo *et al*., 2020). Mekanisme *gate* memungkinkan LSTM untuk memahami konteks kalimat dalam menyeleksi informasi penting, sehingga menghasilkan pemahaman yang lebih baik dan pemrosesan data (Marpaung *et al*., 2021).

Das *et al*. (2021) menggunakan LSTM untuk klasifikasi ujaran kebenncian bahasa Bangla. Hasil penelitian menunjukkan nilai akurasi, presisi, *recall*, dan *f1-score* sebesar 75% menggunakan tujuh kelas yaitu *aggressive comment, poilitical comment, hate speech, ethnical attack, religious hatred, religious comment,* dan *suicidal comment*. Verma *et al*. (2023) menggunakan LSTM untuk klasifikasi ujaran kebencian bahasa Inggris. Hasil penelitian menunjukkan nilai akurasi, presisi, *recall*, dan *f1-score* sebesar 95% namun, hanya menggunakan 2 kelas yaitu *hate* dan *non-hate*.

Penelitian ini menggabungkan teknik augmentasi dan teknik klasifikasi ujaran kebencian pada cuitan *tweet* bahasa Indonesia. Teknik augmentasi yang diajukan

pada penelitian ini adalah *back translation* dan augmentasi BERT untuk menambah jumlah data. Proses augmentasi dilakukan dengan menerjemahkan teks awal ke bahasa Inggris, lalu diproses dengan BERT, dan diterjemahkan kembali ke bahasa Indonesia. Augmentasi BERT digunakan untuk menambah keragaman data tanpa mengubah makna, dan *back translation* digunakan untuk mengatasi korpus yang terbatas.

Klasifikasi teks ujaran kebencian *multi label* dalam bahasa Indonesia menggunakan kombinasi LSTM dan *Transformer*. LSTM digunakan pada bagian awal model untuk menyaring informasi yang kurang penting, sedangkan *Transformer* digunakan untuk memahami hubungan antar kata dalam kalimat secara global. Pendekatan ini bertujuan untuk meningkatkan efisiensi model dalam mengenali berbagai jenis dan tingkatan ujaran kebencian. Hasil pengujian dari metode yang diusulkan akan diukur menggunakan akurasi, presisi, *recall*, dan *f1-score*. Pengukuran keberhasilan model dilakukan untuk melihat performa model dalam melakukan klasifikasi teks ujaran kebencian.

1.2 Perumusan Masalah

Rumusan masalah pada penelitian ini, sebagai berikut:

1. Bagaimana penerapan teknik *back translation* dan augmentasi BERT dalam menambah jumlah data ujaran kebencian pada *tweet* bahasa Indonesia?
2. Bagaimana hasil evaluasi kinerja arsitektur LSTM dan *Transformer* pada klasifikasi teks *dataset* ujaran kebencian *tweet* bahasa Indonesia yang telah

di augmentasi menggunakan ukuran kinerja akurasi, presisi, *recall*, dan *f1-score*?

1.3 Pembatasan Masalah

Pembatasan masalah pada penelitian ini, sebagai berikut:

1. Data yang digunakan pada penelitian ini terdiri dari 44 kelas, yang merupakan kombinasi dari 12 *label* yaitu ujaran kebencian terhadap individu, kelompok, agama, ras, fisik, gender, serta tingkatan keparahannya.
2. Ukuran evaluasi kinerja pada augmentasi dan klasifikasi teks dalam mendeteksi ujaran kebencian pada *tweet* menggunakan akurasi, presisi, *recall*, dan *f1- score.*

1.4 Tujuan

Tujuan penelitian ini, sebagai berikut:

1. Menerapkan teknik *back translation* dan augmentasi BERT dalam menambah jumlah data ujaran kebencian pada *tweet* bahasa Indonesia .
2. Mengetahui hasil evaluasi kinerja arsitektur LSTM dan *Transformer* pada klasifikasi teks *dataset* ujaran kebencian *tweet* bahasa Indonesia yang telah di augmentasi menggunakan ukuran kinerja akurasi, presisi, *recall*, dan *f1-score*.

1.5 Manfaat

Manfaat penelitian ini yaitu dapat diterapkan dalam sistem otomatis untuk mendeteksi ujaran kebencian sesuai dengan jenis dan tingkatannya di media sosial, sehingga dapat mengenali dan mengatasi konten ujaran kebencian dengan efektif.

DAFTAR PUSTAKA

Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information (Switzerland)*, *13*(6), 1–22.

Antypas, D., & Camacho-Collados, J. (2023). Robust hate speech detection in social media: a cross-dataset empirical evaluation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 231–242.

Apicella, A., Donnarumma, F., Isgrò, F., & Prevete, R. (2021). A survey on modern trainable activation functions. *Neural Networks*, *138*(6), 14–32.

Arbaatun, C. N., Nurjanah, D., & Nurrahmi, H. (2022). Hate speech detection on twitter through natural language processing using LSTM model. *Building of Informatics, Technology and Science (BITS)*, *4*(3), 1548–1557.

Aurora, F., Zahra, A., Sibaroni, Y., Sri, &, & Prasetyowati, S. (2023). Classification of Multi-Label of hate speech on twitter Indonesia using LSTM and BiLSTM Method. *JINAV: Journal of Information and Visualization*, *4*(2), 2746–1440.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions. *Computer Science Review*, *38*(9), 100311.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. *ArXiv*, *V1*(7), 1–14.

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, *24*.

Beyhan, F., Çarık, B., Arın, İ., Terzioğlu, A., Yanıkoğlu, B., & Yeniterzi, R. (2022). A Turkish hate speech dataset and detection system. *2022 Language Resources and Evaluation Conference, LREC 2022*, 4177–4185.

Bilal, M., Khan, A., Jan, S., Musa, S., & Ali, S. (2023). Roman urdu hate speech detection using Transformer-based model for cyber security applications. *Sensors*, *23*(8), 1–26.

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553.

Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., & Gu, Q. (2020). Closing the generalization gap of adaptive gradient methods in training deep neural networks. *IJCAI International Joint Conference on Artificial Intelligence*, *2021-Janua*, 3267–3275.

Chotirat, S., & Meesad, P. (2021). Part-of-speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning. *Heliyon*, *7*(10), e08216.

Das, A. K., Al Asif, A., Paul, A., & Hossain, M. N. (2021). Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, *30*(1), 578–591.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using State-of-the-Art NLP models. *Computational Intelligence and Neuroscience*, *2022*(6), 26.

Domor, I., Sun, Y., & Ileberi, E. (2024). Machine Learning with Applications Artificial intelligence and sustainable development in Africa : A comprehensive review. *Machine Learning with Applications*, *18*(July),

Egger, R. (2022). Text Representations and Word Embeddings: Vectorizing Textual Data. *Tourism on the Verge*, *Part F1051*, 335–361.

Elsafoury, F. (2023). Thesis Distillation: Investigating The Impact of Bias in NLP Models on Hate Speech Detection. *BigPicture 2023 - Big Picture Workshop, Proceedings*, *June*, 53–65.

Ezhilarasi, S., & Maheswari, D. P. U. (2021). Designing the Neural Model for POS Tag classification and prediction of words from ancient stone inscription script. *Int. J. of Aquatic Science*, *12*(3), 1718–1728.

Fazil, M., Khan, S., Albahlal, B. M., Alotaibi, R. M., Siddiqui, T., & Shah, M. A. (2023). Attentional Multi-Channel Convolution with Bidirectional LSTM cell toward hate speech prediction. *IEEE Access*, *11*(2), 16801–16811.

Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). Survey on Text Classification Algorithms: from text to predictions. *Information (Switzerland)*, *13*(2), 1–39.

Hana, K. M., Adiwijaya, Al Faraby, S., & Bramantoro, A. (2020). Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines. *2020 International Conference on Data Science and Its Applications, ICoDSA 2020*.

Hernández, A., & Amigó, J. M. (2021). Attention mechanisms and their

applications to complex systems. *Entropy*, *23*(3), 1–18.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian twitter. *2019 Association for Computational Linguistics (ACL)*, 46–57.

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. *Heliyon*, *9*(8), e18647.

Kapil, P., & Ekbal, A. (2022). A Transformer based Multi-Task Learning Approach Leveraging Translated and Transliterated Data to Hate Speech Detection in Hindi. *Computer Science & Informasion Technology (CS & IT)*, 191–207.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*, *54*(10), 1–30.

Kolesnichenko, L., Velldal, E., & Øvrelid, L. (2023). Word substitution with Masked Language Models as data augmentation for sentiment analysis. *RESOURCEFUL. 2023 - Workshop on Resources and Representations for Under-Resourced Languages and Domains, Proceedings of the 2nd*, *5*(task 10), 42–47.

Kovaleva, O., Romanov, A., Rogers, A., & Rumshisky, A. (2019). Revealing the dark secrets of Bert. *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*, *8*(2018), 4365–4374.

Kurniawan, S., & Budi, I. (2020). Indonesian tweets hate speech target classification using machine learning. *2020 5th International Conference on Informatics and Computing, ICIC 2020*, 1–5.

Li, J., Wang, X., Tu, Z., & Lyu, M. R. (2021). On the diversity of multi-head attention. *Neurocomputing*, *454*, 14–24.

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 1–10.

Liu, L., Liu, X., Gao, J., Chen, W., & Han, J. (2020). Understanding the difficulty of training Transformers. *Conference on Empirical Methods in Neural Language Processing*, 5747–5763.

Malik, J. S., Qiao, H., Pang, G., & Hengel, A. van den. (2022). Deep Learning for Hate Speech Detection: A Comparative Study. *International Journal of Data Science and Analytics*, *09517v2*(12), 1–18.

Marpaung, A., Rismala, R., & Nurrahmi, H. (2021). Hate speech detection in

Indonesian twitter texts using Bidirectional Gated Recurrent Unit. *KST 2021 - 2021 13th International Conference Knowledge and Smart Technology*, 186–190.

Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable AI. *Algorithms*, *15*(8), 291.

Mohiuddin, K., Welke, P., Alam, M. A., Martin, M., Alam, M. M., Lehmann, J., & Vahdati, S. (2023). Retention Is All You Need. *International Conference on Information and Knowledge Management, Proceedings*, *Nips*, 4752–4758.

Mostafa, G., Ahmed, I., & Junayed, M. S. (2021). Investigation of different Machine Learning Algorithms to determine human sentiment using twitter data. *International Journal of Information Technology and Computer Science*, *13*(2), 38–48.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. *Studies in Computational Intelligence*, *881 SCI*, 928–940.

Ohiri, E. (2024). *Feedforward neural networks: everything you need to know*. CudaCompute Blog. https://www.cudocompute.com/blog/feedforward-neural-networks-everything-you-need-to-know?

Or, C., Quintana, C., & Zilio, L. (2018). Challenges in Translation of Emotions in Multilingual User-Generated Content : Twitter as a Case Study. *ArXiv*, *1*(6).

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors (Switzerland)*, *19*(21), 1–37.

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, *VI*(10), 37–63.

Qiu, D., & Yang, B. (2022). Text summarization based on multi-head self-attention mechanism and pointer network. *Complex and Intelligent Systems*, *8*(1), 555–567.

Rahman, M. M., Balakrishnan, D., Murthy, D., Kutlu, M., & Lease, M. (2021). An information retrieval approach to building datasets for hate hpeech detection. *ArXiv Computer Science*, *11*(NeurIPS 2021), 1–28.

Rongbo, Z., XU, Z., & IWAIHARA, M. (2020). Multi-Label Text Classification Using Only Label Names. *DEIM Forum*, 1–10.

Sabty, C., Omar, I., Wasfalla, F., Islam, M., & Abdennadher, S. (2021). Data Augmentation Techniques on Arabic Data for Named Entity Recognition.

Procedia CIRP, 292–299.

Saeidi, M., Samuel, S. B., Milios, E., Zeh, N., & Berton, L. (2020). Categorizing Online Harassment on Twitter. In *Communications in Computer and Information Science: Vol. 1168 CCIS*.

Salau, A., & Yesufu, T. K. (2020). *Recent Trends in Image and Signal Processing in Computer Vision*.

Sarkar, D., Zampieri, M., Ranasinghe, T., & Ororbia, A. (2021). FBERT: A Neural Transformer for Identifying Offensive Content. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, *1*(9), 1792–1798.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing.

Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., & Takalikar, M. S. (2023). Language augmentation approach for code-mixed text classification. *Natural Language Processing Journal*, *5*(November), 100042.

Tra, V., Duong, B. P., & Kim, J. M. (2019). Improving diagnostic performance of a power transformer using an adaptive over-sampling method for imbalanced data. *IEEE Transactions on Dielectrics and Electrical Insulation*, *26*(4), 1325–1333.

Vasiu, M. A., & Potolea, R. (2020). Enhancing Tokenization by Embedding Romanian Language Specific Morphology. *Proceedings - 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing, ICCP 2020*, 243–250.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Verma, A., Singh, A., Bihari, A., Tripathi, S., Agrawal, S., Pandey, S. K., & Verma, S. (2023). Identification of Hate Speech on Social Media using LSTM. *GMSARN International Journal*, *17*(4), 468–474.

Wang, C., Nulty, P., & Lillis, D. (2020). A Comparative Study on Word Embeddings in Deep Learning for Text Classification. *ACM International Conference Proceeding Series*, 37–46.

Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2019). Conditional BERT contextual augmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11539 LNCS*(12), 84–95.

Xia, M., Kong, X., Anastasopoulos, A., & Neubig, G. (2020). Generalized data augmentation for low-resource translation. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 5786–5796.

Yi, D., Ahn, J., & Ji, S. (2020). An effective optimization method for machine learning based on ADAM. *Applied Sciences (Switzerland)*, *10*(3).

Zhang, Y., Wen, J., Yang, G., He, Z., & Wang, J. (2019). Path loss prediction based on machine learning: Principle, method, and data expansion. *Applied Sciences (Switzerland)*, *9*(9).