Jurnal Komputer dan Informatika
Vol 18 No 1, April 2023: hlm 31-37

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

KLASIFIKASI TOKSISITAS KOMENTAR DENGAN ALGORITMA NAIVE BAYES DAN *DECISION TREE*

David Ciang
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara,
Jln. Letjen S.Parman No. 1,Jakarta, 11440, Indonesia
e-mail: *david.535200011@stu.untar.ac.id*

ABSTRAK

Studi ini bertujuan mengembangkan model klasifikasi toksisitas komentar menggunakan algoritma Naive Bayes dan *Decision Tree*, khususnya dalam konteks lingkungan daring. Dataset terdiri dari komentar-komentar daring yang melibatkan proses preprocessing, termasuk pembersihan dan normalisasi teks, serta ekstraksi fitur dengan metode seperti TF-IDF. Model klasifikasi Naive Bayes dan *Decision Tree* dilatih menggunakan dataset tersebut, dan evaluasi kinerja model dilakukan dengan metrik standar seperti akurasi, presisi, *recall*, dan *F1-Score*. Selain itu, analisis perbandingan antara Naive Bayes dan *Decision Tree* dilakukan, fokus pada konteks daring. Analisis ini bertujuan memberikan wawasan terkait keefektifan keduanya dalam mengidentifikasi toksisitas komentar dalam lingkungan daring. Temuan studi ini dapat dijadikan dasar untuk pengembangan solusi moderasi konten yang mampu beradaptasi dengan dinamika interaksi manusia di dunia daring.

Hasil penelitian ini memiliki implikasi penting dalam membangun sistem moderasi konten yang lebih efisien dan efektif dalam lingkungan daring. Dengan fokus pada konteks daring, penelitian ini memberikan kontribusi berharga terhadap pemahaman tentang performa algoritma klasifikasi dalam menghadapi toksisitas komentar dalam interaksi online. Dengan demikian, temuan studi ini dapat membantu meningkatkan keamanan dan kenyamanan pengguna di lingkungan daring melalui pengembangan solusi moderasi konten yang lebih canggih.

Kata kunci: Algoritma Naïve Bayes, Algoritma *Decision Tree*, TF-IDF, Akurasi, Toksisitas

ABSTRACT

This study aims to develop a toxicity comment classification model using Naive Bayes and Decision Tree algorithms, specifically in the context of the online environment. The dataset consists of online comments, involving preprocessing steps such as text cleaning, normalization, and feature extraction using methods like TF-IDF. The Naive Bayes and Decision Tree classification models are trained on this dataset, and their performance is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, a comparative analysis between Naive Bayes and Decision Tree is conducted, focusing on the online context. This analysis aims to provide insights into their effectiveness in identifying toxicity in online comments. The findings of this study serve as a foundation for developing content moderation solutions that can adapt to the dynamic nature of human interactions in the online world.

The results of this research have significant implications for building more efficient and effective content moderation systems in the online environment. By concentrating on the online context, the study makes a valuable contribution to understanding the performance of classification algorithms in addressing toxicity in online interactions. Consequently, the study's findings can help enhance user safety and comfort in the online environment through the development of more sophisticated content moderation solutions.

Keywords: *Naïve Bayes Algorithm, Decision Tree Algorithm, Accuracy, TF-IDF, Toxicity.*

1 PENDAHULUAN

Dengan pesatnya pertumbuhan komunikasi daring dan luring, meningkatnya partisipasi pengguna dalam berbagai platform telah memberikan kontribusi positif pada akses informasi dan interaksi manusia. Namun, perkembangan ini juga menghadirkan tantangan baru, terutama dalam

mengatasi komentar toksik yang dapat merusak lingkungan diskusi dan interaksi manusia. Toksisitas komentar, baik dalam lingkungan daring maupun luring, dapat menciptakan atmosfer yang tidak sehat, mengancam keamanan, dan menghambat dialog yang konstruktif. Pentingnya mengidentifikasi dan mengatasi komentar toksik menciptakan urgensi untuk pengembangan model klasifikasi yang efektif dan adaptif. Perbedaan konteks antara interaksi daring dan luring menambah kompleksitas permasalahan, memerlukan pendekatan yang holistik untuk memahami dan mengatasi toksisitas komentar di kedua situasi. Penelitian ini bertujuan untuk mengembangkan model klasifikasi toksisitas komentar menggunakan algoritma Naive Bayes dan *Decision Tree*. Pemilihan kedua algoritma tersebut didasarkan pada keunggulan mereka dalam mengolah data teks dengan kompleksitas bahasa. Dengan memanfaatkan dataset yang mencakup kedua konteks, penelitian ini diharapkan memberikan pemahaman yang mendalam terhadap efektivitas algoritma dalam menghadapi toksisitas komentar dalam berbagai situasi komunikasi.

Pernyataan masalah dalam penelitian ini adalah "Bagaimana mengembangkan model klasifikasi toksisitas komentar menggunakan algoritma Naive Bayes dan *Decision Tree* untuk mengatasi tantangan dalam lingkungan daring dan luring?". Justifikasi penelitian ini terletak pada kebutuhan mendesak untuk solusi moderasi konten yang adaptif dan efektif dalam mengatasi kompleksitas toksisitas komentar dalam kedua konteks tersebut. Tujuan utama penelitian ini adalah menghasilkan model klasifikasi toksisitas komentar yang efektif untuk lingkungan daring dan luring menggunakan pendekatan algoritma Naive Bayes dan *Decision Tree*. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem moderasi konten yang lebih cerdas dan adaptif. Selain itu, penelitian ini dapat memberikan wawasan yang berguna bagi pihak-pihak terkait dalam merancang kebijakan dan alat moderasi yang lebih responsif terhadap dinamika interaksi manusia dalam berbagai konteks komunikasi.

2 TINJAUAN LITERATUR

Berdasarkan tinjauan literatur yang dilakukan, hipotesis penelitian dirumuskan dengan mengajukan dua pernyataan. Hipotesis nol menyatakan bahwa tidak ada perbedaan signifikan dalam kinerja klasifikasi toksisitas komentar antara algoritma Naive Bayes dan *Decision Tree* dalam konteks daring. Sebaliknya, hipotesis alternatif menyatakan bahwa terdapat perbedaan signifikan dalam kinerja klasifikasi toksisitas komentar antara kedua algoritma dalam konteks daring. Tinjauan literatur juga menunjukkan bahwa kompleksitas toksisitas komentar dalam konteks daring menjadi perhatian utama, dengan sedikit penelitian yang secara khusus mengeksplorasi klasifikasi toksisitas komentar dalam interaksi langsung. Algoritma klasifikasi teks, seperti Naive Bayes dan *Decision Tree*, terbukti efektif, namun perbandingan mendalam kinerja keduanya untuk tugas klasifikasi toksisitas komentar masih perlu dieksplorasi lebih lanjut. Dengan merinci klasifikasi toksisitas komentar menggunakan kedua algoritma dan mempertimbangkan representasi vektor yang optimal, penelitian ini bertujuan memberikan pemahaman mendalam dan solusi terfokus terhadap tantangan klasifikasi toksisitas komentar dalam kedua konteks tersebut. Dengan menguji hipotesis ini, diharapkan penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang keefektifan masing-masing algoritma dalam mengatasi permasalahan klasifikasi toksisitas komentar dalam kedua konteks tersebut.

3 METODE PENELITIAN

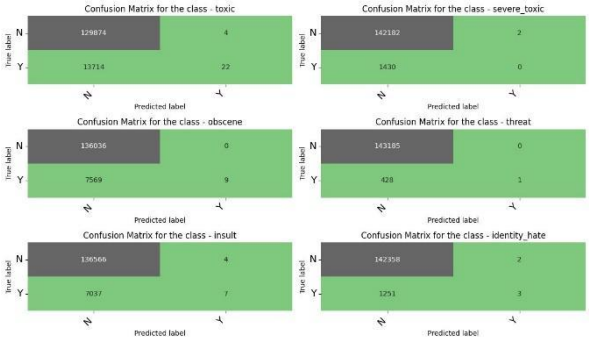
Metode penelitian ini difokuskan pada klasifikasi toksisitas komentar dengan memanfaatkan algoritma Naive Bayes dan *Decision Tree* dalam konteks daring. Data penelitian diperoleh dari komentar-komentar yang terdapat di platform media sosial daring. Alat penelitian mencakup perangkat lunak pemrosesan bahasa alami, algoritma klasifikasi teks, dan metode representasi vektor, terutama menggunakan teknik TF-IDF. Pendekatan penelitian bersifat kuantitatif dengan rancangan

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

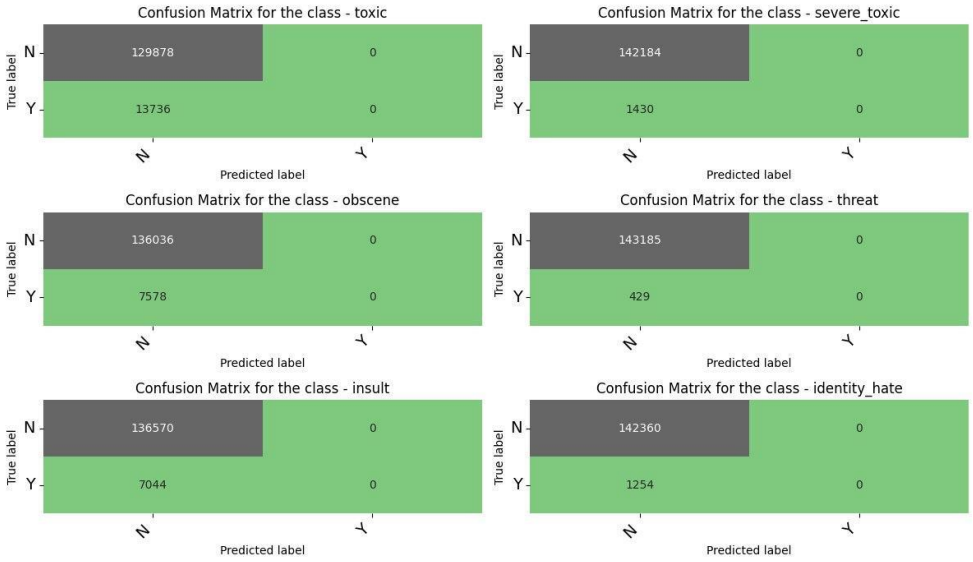
kegiatan yang melibatkan pengumpulan data, pengolahan data, dan evaluasi kinerja algoritma. Ruang lingkup penelitian mencakup variasi komentar dari berbagai topik untuk memastikan keragaman data. Bahan utama penelitian terdiri dari korpus komentar yang diperoleh dari sumber-sumber daring yang relevan dengan topik penelitian. Alat utama melibatkan perangkat lunak pemrosesan bahasa alami, pengembangan model klasifikasi dengan algoritma Naive Bayes dan *Decision Tree*, serta perangkat lunak analisis hasil. Teknik pengumpulan data dilakukan dengan mengambil komentar-komentar dari platform media sosial daring yang sesuai dengan tujuan penelitian. Definisi operasional variabel penelitian mencakup parameter evaluasi kinerja seperti akurasi, presisi, *recall*, dan *F1-Score*. Analisis data melibatkan perbandingan kinerja antara algoritma Naive Bayes dan *Decision Tree* dalam konteks daring, dengan mempertimbangkan representasi vektor yang optimal. Dengan pendekatan ini, penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam tentang efektivitas masing-masing algoritma dalam menangani permasalahan klasifikasi toksisitas komentar dalam lingkungan daring.)

4 HASIL DAN PEMBAHASAN

Setelah melakukan percobaaan dan melakukan optimalisasi kode program. Dimana pengujian dilakukan dengan membagi data training menjadi 2 jenis percobaan yaitu 10% data training dan 90% data training. Keduanya diuji ke dalam algoritma yang telah dibuat dan menghasilkan data yang menunjukkan bahwa algoritma *Decision Tree* memiliki kemampuan yang lebih baik dalam melakukan klasifikasi toksisitas komentar. Berikut adalah hasil training algoritma *Decision Tree* dengan menggunakan 10% Data Training. Menurut hasil algortitma ditermukan bahwa komentar yang dikategorikan sebagai toksik, toksik parah (*severe toxic*), cabul (*obscene*), ancaman (*threat*), hinaan (*insult*), dan rasis (*identity hate*) terdiri dari 13736 komentar di kategorikan sebagai toksik, 1430 komentar dikategorikan sebagai toksis parah, 7578 komentar dikategorikan sebagai cabul, 429 komentar dikategorikan sebagai ancaman, 7044 komentar dikategorikan sebagai hinaan, 1254 komentar dikategorikan sebagai rasis dengan precision berada di ambang 0-1 dimana 0 dimiliki oleh data toksik parah dan 1 dimiliki oleh data cabul dan ancaman , *recall* yang berada pada ambang 0 - 0.0023 dengan 0 dimiliki data toksik parah dan 0.0023 dimiliki oleh ancaman kemudian *F1-Score* berada pada ambang 0 – 0.0046 dengan 0 dimiliki oleh data toksik parah dan 0.0046 dimiliki oleh data ancaman.

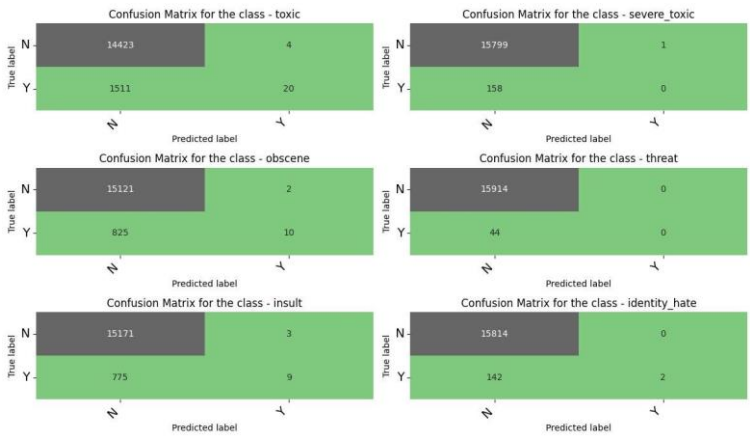


	precision	recall	f1-score	support
toxic	0.846154	0.001602	0.003197	13736
severe_toxic	0.000000	0.000000	0.000000	1430
obscene	1.000000	0.001188	0.002372	7578
threat	1.000000	0.002331	0.004651	429
insult	0.636364	0.000994	0.001984	7044
identity_hate	0.600000	0.002392	0.004766	1254
micro avg	0.777778	0.001335	0.002665	31471
macro avg	0.680420	0.001418	0.002828	31471
weighted avg	0.790083	0.001335	0.002664	31471
samples avg	0.000144	0.000139	0.000137	31471



	precision	recall	f1-score	support
toxic	0.000000	0.000000	0.000000	13736
severe_toxic	0.000000	0.000000	0.000000	1430
obscene	0.000000	0.000000	0.000000	7578
threat	0.000000	0.000000	0.000000	429
insult	0.000000	0.000000	0.000000	7044
identity_hate	0.000000	0.000000	0.000000	1254
micro avg	0.000000	0.000000	0.000000	31471
macro avg	0.000000	0.000000	0.000000	31471
weighted avg	0.000000	0.000000	0.000000	31471
samples avg	0.000000	0.000000	0.000000	31471

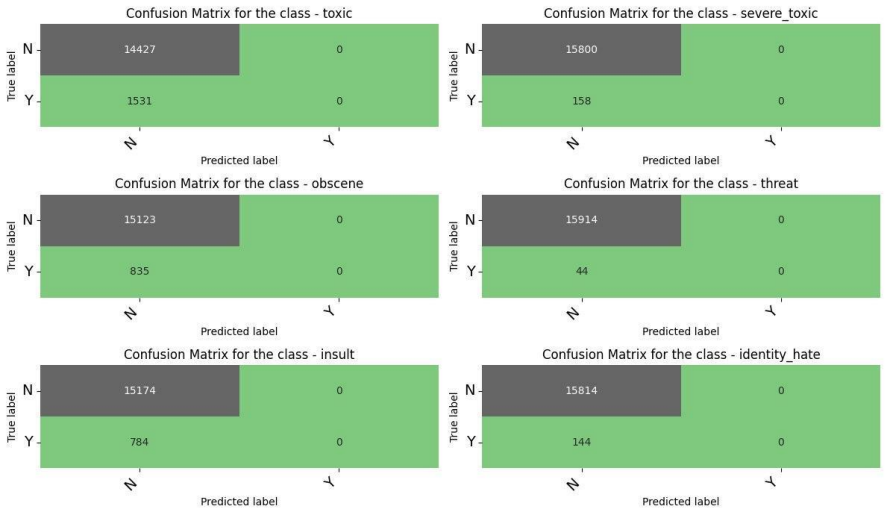
Berikut merupakan hasil training algoritma Naïve Bayes menggunakan 10% data training. Menurut hasil algortitma ditermukan bahwa komentar yang dikategorikan sebagai toksik, toksik parah (*severe toxic*), cabul (*obscene*), ancaman (*threat*), hinaan (*insult*), dan rasis (*identity hate*) terdiri dari 13736 komentar di kategorikan sebagai toksik, 1430 komentar dikategorikan sebagai toksis parah, 7578 komentar dikategorikan sebagai cabul, 429 komentar dikategorikan sebagai ancaman, 7044 komentar dikategorikan sebagai hinaan, 1254 komentar dikategorikan sebagai rasis yang terdiri dari 0 *precision* pada semua kategori komentar kemudian 0 *recall* pada semua kategori komentar, 0 *F1-Score* pada semua kategori komentar



ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

	precision	recall	f1-score	support
toxic	0.833333	0.013063	0.025723	1531
severe_toxic	0.000000	0.000000	0.000000	158
obscene	0.833333	0.011976	0.023613	835
threat	0.000000	0.000000	0.000000	44
insult	0.750000	0.011480	0.022613	784
identity_hate	1.000000	0.013889	0.027397	144
micro avg	0.803922	0.011728	0.023118	3496
macro avg	0.569444	0.008401	0.016558	3496
weighted avg	0.773360	0.011728	0.023104	3496
samples avg	0.001196	0.001058	0.001069	3496

Berikut adalah hasil training algoritma *Decision Tree* dengan menggunakan 90% Data Training. Menurut hasil algortitma ditermukan bahwa komentar yang dikategorikan sebagai toksik, toksik parah (*severe toxic*), cabul (*obscene*), ancaman (*threat*), hinaan (*insult*), dan rasis (*identity hate*) terdiri dari 1531 komentar di kategorikan sebagai toksik, 158 komentar dikategorikan sebagai toksis parah, 835 komentar dikategorikan sebagai cabul, 44 komentar dikategorikan sebagai ancaman, 784 komentar dikategorikan sebagai hinaan, 144 komentar dikategorikan sebagai rasis dengan precision berada di ambang 0-1dimana 0 dimiliki oleh data toksik parah dan data ancaman sedangkan 1 dimiliki oleh data rasis, *recall* yang berada pada ambang 0 - 0.013 dengan 0 dimiliki data toksik parah dan data ancaman sedangkan 0.013 dimiliki oleh toksik dan rasis kemudian *F1-Score* berada pada ambang 0 – 0.027 dengan 0 dimiliki oleh data toksik parah dan data ancaman sedangkan 0.027 dimiliki oleh data rasis.



	precision	recall	f1-score	support
toxic	0.000000	0.000000	0.000000	1531
severe_toxic	0.000000	0.000000	0.000000	158
obscene	0.000000	0.000000	0.000000	835
threat	0.000000	0.000000	0.000000	44
insult	0.000000	0.000000	0.000000	784
identity_hate	0.000000	0.000000	0.000000	144
micro avg	0.000000	0.000000	0.000000	3496
macro avg	0.000000	0.000000	0.000000	3496
weighted avg	0.000000	0.000000	0.000000	3496
samples avg	0.000000	0.000000	0.000000	3496

Berikut adalah hasil dari algoritma Naïve bayes yang menggunakan 90% data *training*. Menurut hasil algortitma ditermukan bahwa komentar yang dikategorikan sebagai toksik, toksik parah (*severe toxic*), cabul (*obscene*), ancaman (*threat*), hinaan (*insult*), dan rasis (*identity hate*) terdiri dari 1531 komentar di kategorikan sebagai toksik, 158 komentar dikategorikan sebagai toksis parah, 835 komentar dikategorikan sebagai cabul, 44 komentar dikategorikan sebagai ancaman, 784 komentar dikategorikan sebagai hinaan, 144 komentar dikategorikan sebagai rasis dengan *precision, recall, F1-Score* berada pada nilai 0 pada semua data.

5 KESIMPULAN

Kesimpulan yang dapat diambil dari penelitian ini adalah penggunaan algoritma *Decision Tree* dalam melakukan klasifikasi toksisitas komentar sangat disarankan dikarenakan dapat memberikan hasil yang signifikan apabila dibandingkan dengan data yang didapat menggunakan algoritma Naïve Bayes. Walaupun algoritma Naïve Bayes dapat membaca data dan menghasilkan data TP (*True Positive*) dalam confusion matrix yang lebih besar dari pada hasil TP yang dihasilkan oleh confusion matrix algoritma *Decision Tree*.

UCAPAN TERIMA KASIH

Ingin saya sampaikan rasa terima kasih yang tulus kepada semua pihak yang telah turut serta dalam perjalanan ini. Pertama-tama, kepada Tuhan Yang Maha Esa, yang memberikan petunjuk dan berkat sehingga jurnal ini dapat terwujud. Kepada orang tua, terima kasih atas dukungan tak terhingga dan waktu produktif yang diberikan dalam lingkungan rumah. Tidak lupa kepada teman-teman yang senantiasa memberikan semangat dan saran perbaikan baik untuk jurnal maupun kode program. Keterlibatan dan kontribusi kalian memberikan nilai tambah yang luar biasa. Spesial terima kasih kepada tim ChatGPT atas dedikasi mereka dalam mengembangkan kemampuan model bahasa. Platform yang disediakan tidak hanya mendorong inovasi, tetapi juga membantu dalam memperbaiki dan mengatasi berbagai tantangan seperti permasalahan kode program, memberi saran penulisan untuk beberapa bagian dalam jurnal. Wawasan dan panduan yang diberikan oleh tim ChatGPT telah menjadi kunci keberhasilan proyek ini.

Saya juga sangat bersyukur atas kerjasama kolaboratif dari semua individu yang dengan tulus berbagi pengetahuan, waktu, dan masukan berharga. Kontribusi kalian menjadi faktor utama dalam mengidentifikasi area yang perlu diperbaiki dan mengimplementasikan solusi yang efektif.Perjalanan kolaboratif ini memberikan pengalaman berharga, dan saya sangat menghargai dampak positif yang diberikan. Sekali lagi, terima kasih kepada semua pihak yang terlibat atas dedikasi dan dukungan tanpa henti. Saya berharap dapat terus berkolaborasi dan mencapai kemajuan lebih lanjut di masa depan.

DAFTAR PUSTAKA

[1] D. Jurafsky dan J. H. Martin, *Speech and Language Processing*, 2022.

[2]. C. D. Manning, P. Raghavan dan H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2022.

[3]. J. Chen, L. Song, W. Li, Y. Zhang dan X. Cheng, "Exploring Sentiment in Social Media: A Comprehensive Survey.," *Knowledge-Based Systems*, vol. 198, p. 105947, 2023.

[4] B. Pang dan L. Lee, "Opinion Mining and Sentiment Analysis: Foundations and Trends," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2019.

[5]. F. Sebastina, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys (CSUR)*, vol. 34, no. 1, pp. 1-47, 2017.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

[6]. A. Srivasta dan V. Singh, "A Comprehensive Review on Text Mining using Novel Methods," *Procedia Computer Science,* vol. 165, pp. 197-204, 2023
[7]. S. Tan, X. Cheng dan Y. Wang, "Feature Engineering and Selection for Text Classification: A Review," *Data and Knowledge Engineering,* vol. 100, pp. 13-21, 2021.
[8]. H. Witten, E. Frank dan M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2021.
[9]. S. Kim, *"Mining Twitter Data with Python* (Part 1: Collecting Data)," 2018.
[10]. J. Saldaña, *The Coding Manual for Qualitative Researchers*, SAGE Publications, 2017.
[11]. I. Rish, "An Empirical Study of the Naive Bayes Classifier," dalam *IJCAI 2011 Workshop on Empirical Methods in Artificial Intelligence*, 2011.
[12]. J. S. R. Pennington dan C. D. Manning, "GloVe: Global Vectors for Word Representation," dalam *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.
[13]. S. Hochreiter dan J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 2017.
[14]. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado dan J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," dalam *Advances in Neural Information Processing Systems*, 2018.
[15]. P. Shrestha, A. Mahmood dan E. Yafi, "A Comprehensive Survey of Machine Learning Techniques in Sentiment Analysis," *Information Processing & Management*, vol. 56, no. 5, pp. 1794-1818, 2023.
[16]. Y. Yang dan J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *dalam Proceedings of the Fourteenth International Conference on Machine Learning,* 2017.
[17]. S. R. Makhija dan P. Srinivasan, "Text Classification Using Deep Learning Models: A Comprehensive Review," *Journal of King Saud University - Computer and Information Sciences,* 2022.
[18]. Y. Zhang dan B. Wallace, "A Survey of Emerging Trends in Sentiment Analysis in Social Media," *Journal of Artificial Intelligence Research*, vol. 71, pp. 933-993, 2021
[19]. C. E. dan W. B., "Jumping NLP Curves: A Review of Natural Language Processing Research," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48-57, 2019.
[20]. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2016.