Prediksi Ujaran Kebencian Berbasis Text Pada Sosial Media Menggunakan Metode Neural Network

Prediction of Text Based Hate Speech on Socials Media Using the Naural Network

Kristiawan Nugroho*[1], Endang Tjahjaningsih [2], Lie Liana [2], Raden Mohamad Herdian Bhakti[3]
[1]*Program Studi Magister Teknologi Informasi, Fakultas Teknologi Informasi dan Industri, Universitas Stikubank*
[2]*Program Studi Magister Manajemen, Fakultas Ekonomi dan Bisnis , Universitas Stikubank*
[3]*Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhadi Setiabud*
*email: *[1]kristiawan@edu.unisbank,ac,id, [2]naning@edu.unisbank.ac.id ,[2]lieliana@edu.unisbank.ac.id ,[3]herdian.bhakti@umus.ac.id*

Abstrak

Saat ini teknologi informasi telah membantu dalam berbagai bentuk kehidupan manusia. Mereka dapat berkomunikasi satu sama lain melalui berbagai media elektronik, termasuk menggunakan media sosial. Jumlah pengguna media sosial semakin meningkat dari tahun ke tahun di Indonesia. Namun perkembangan penggunaan media sosial juga menimbulkan berbagai permasalahan, termasuk ujaran kebencian yang pada akhirnya akan menimbulkan akibat hukum. Berbagai cara ditempuh untuk membatasi berkembangnya ujaran kebencian, antara lain dengan memblokir pengguna yang menulis ujaran kebencian di aplikasi media sosial. Pembatasan penggunaan media sosial untuk ujaran kebencian dapat lebih optimal dilakukan dengan mendeteksi kata-kata berbasis teks yang berpotensi menjadi ujaran kebencian. Penelitian ini menggunakan metode Neural Network (NN) untuk memprediksi kata-kata yang mengandung ujaran kebencian di media sosial dengan tingkat akurasi 73% lebih baik dibandingkan dengan metode lain seperti Decission Tree dan K-Nearest Neighbor (KNN) yang hanya mencapai tingkat akurasi 68,5 %..

Kata kunci*—Hate Speech,Sosial Media,Prediksi,Neural Network*

Abstract

Currently information technology has helped in various forms of human life. They can communicate with each other through various electronic media, including using social media. The number of social media users is increasing from year to year in Indonesia. However, the development of the use of social media has also resulted in various problems, including hate speech, which will eventually lead to legal consequences. Various methods have been taken to limit the development of hate speech, including by blocking users who write hate speech on social media applications. Limiting the use of social media for hate speech can be more optimally carried out by detecting text-based words that have the potential to become hate speech. This study uses the Neural Network (NN) method to predict words that contain hatespeech on social media with an accuracy rate of 73% better than other methods such as Decission Tree and K-Nearest Neighbor (KNN) which only achieve an accuracy rate of 68.5 %.

Keywords— *Hate Speech, Social Media, Prediction,Neural Network*

PENDAHULUAN

Dunia digital manusia semakin berkembang sampai saat ini, antara lain pada bidang sosial media di Indonesia. Sosial media merupakan aplikasi yang berkembang dan sangat disukai di Indonesia, Menurut data yang disampaikan oleh Hootsuite pada Indonesian Digital Report 2021, Pengguna sosial media di Indonesia adalah sekitar 170 juta atau 61.8% dari total populasi penduduk.

Informasi Artikel:
Submitted: Februari 2023, **Accepted:** Mei 2023, **Published:** Mei 2023
ISSN: 2685-4902 (media online), Website: http://jurnal.umus.ac.id/index.php/intech

Menurut sumber yang sama ada 5 macam sosial media yang sering dipergunakan di Indonesia yaitu Youtube,Whatapp,Instagram, Facebook dan Twitter. Sedangkan menurut Harahap[1] penggunaan sosial media di Indonesia didominasi oleh beberapa aplikasi yaitu facebook,Whatapp,Instagram dan Telegram. Namun menurut Juditha[2] selain sosial media tersebut diatas, Pengguna Twitter sudah lebih dari 50 Juta dan masih berkembang sampai saat ini

Salah satu permasalahan yang sering ditemui pada penggunaan sosial media adalah ujaran kasar maupun ujaran kebencian.Permasalahan ini seringkali membawa seseorang berhadapan dengan hokum untuk mempertanggung jawabkan perbuatannya. Hatespeech merupakan tidank pidana berupa penghinaan yang bisa membuat rasa benci maupun permusuhan antar individu maupun masyarakat[3] yang saat ini sering dilakukan melalui sosial media.

Permasalahan ini merupakan fenomena serius yang dihadapi pemerintah dan harus segera diselesaikan karena bisa mengakibatkan perselisihan dan berpotensi menyebabkan kekacauan dalam masyarakat.Salah satu metode yang dipergunakan adalah dengan menggunakan Data Mining yang merupakan pendekatan yang sering dipergunakan dalam melakukan prediksi kata-kata yang berjenis hatespeech pada social media. Berbagai penelitian sudah dilakukan untuk mendeteksi adanya hatespeech pada sosial media antara lain dengan metode Naïve Bayes seperti riset yang dilakukan oleh Hakiem[4] dengan Naïve Bayes berbasis N-Gram maupun penelitian Willianto[5] tentang klasifikasi ujaran kebencian di Facebook.

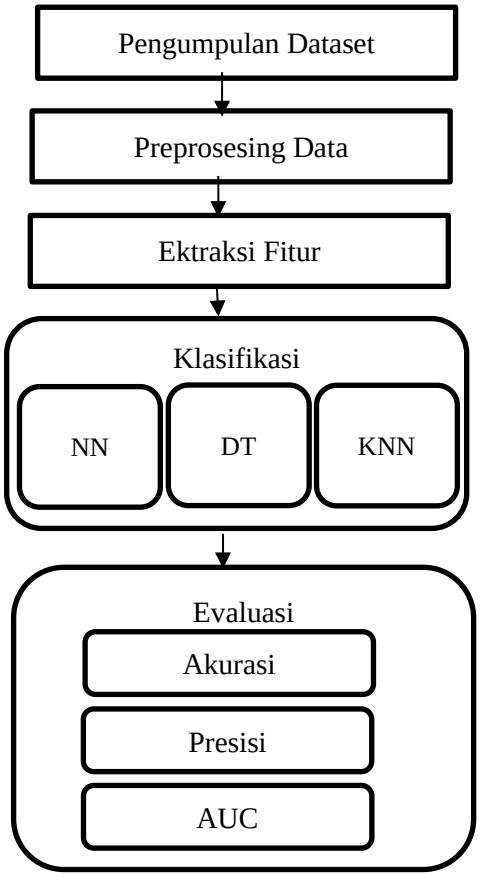
Metode lain yang dipergunakan dalam mengklasifikasikan Hatespeech adalah Support Vector Machine (SVM) seperti riset yang dilakukan oleh Rahman[6] pada Twitter, Ulfah[7] yang melakukan analisis hatespeech pada portal berita online dan Budi[8] yang mengkombinasikan antara Word2Vec dengan SVM. Metode Decision Tree juga dipergunakan dalam mendeteksi ujaran kebencian seperti pada penelitian yang dilakukan oleh Ihsan[9] yang mencapai tingkat akurasi sebesar 70.48% untuk mendeteksi ujaran kebencian pada Twitter. Selain itu Malmasi[10] juga menggunakan Decision Tree dikombinasikan dengan N-Gram-based Word Scoring (TF-IDF) untuk mendeteksi hatespeech pada Twitter.

Berbagai macam metode pada Datamining seperti Naïve Bayes,SVM maupun Decision Tree telah dipergunakan untuk mendeteksi ujaran kebencian pada sosial media, Namun keterbatasan data yang dianalisa dan proses preprocessing data yang kurang tepat sering menjadikan hasil akurasi metode yang dipergunakan menjadi kurang optimal. Penelitian ini mempergunakan metode Neural Network (NN) yang merupakan metode yang robust dan menghasilkan tingkat akurasi yang tinggi dalam berbagai pemrosesan data bertipe teks,gambar, suara maupun video. Pendekatan NN akan dikombinasikan dengan pengolahan preprocessing data dengan cara menghapus semua data yang kosong serta memmpergunakan Teknik normalisasi fitur dengan center by mean yang diharapkan akan meningkatkan kinerja metode NN dalam mendeteksi ujaran teks ujaran kebencian sehingga akan meningkatkan akurasi, presisi maupun recall pada penelitian ini.

METODE PENELITIAN

Metode penelitian merupakan hal yang sangat penting dalam suatu penelitian untuk menjelaskan ururtan sistematis setiap tahapan yang dipergunakan para peneliti dalam suatu penelitian. Menurut Sugiyono[11] metode penelitian pada dasarnya merupakan cara ilmiah untuk mendapatkan data dengan tujuan dan kegunaan tertentu. Tahapan penelitian pada paper ini menggunakan teknik eksperimental dimana setiap tahapan disusun secara sistematis sesuai kaidah ilmiah yang semestinya. Pada paper ini akan disampaikan beberapa tahapan yang akan dipergunakan dalam penelitian mengenai deteksi ujaran kebencian pada sosial media terutama Twitter. Tahapan penelitian tersebut dapat dilihat pada gambar 1 sebagai berikut :

ISSN: 2685-4902 (online)



Gambar 1 Tahapan penelitian

2.1 Pengumpulan Dataset

Penelitian ini menggunakan dataset Indonesian abusive and hastespeech pada Twitter[12] yang diambil dari https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text?resource=download.
Dataset ini terdiri dari 12 atribut dimana atribut abusive merupakan label dengan keterangan 1 untuk data yang berisi ujaran kasar dan 0 untuk data yang berisi bukan ujaran kasar. Selain itu dataset ini terdiri dari 13.169 data

2.2 Preprosesing Data

Data preprocessing merupakan tahapan yang penting dalam mempersiapkan data sebelum dilakukan proses klasifikasi. Pada penelitian ini tahapan preprocessing yang dilakukan adalah dengan menggunakan :

2.2.1 Impute Missing Values

Merupakan tahapan preprocessing dengan cara melakukan penghapusan pada beberapa baris data yang kosong sehingga semua data dipastikan ada baris datanya.

2.2.2 Normalize Feature

Tahapan ini merupakan proses yang dilakukan untuk menghilangkan efek sejumlah fitur kuantitatif yang diukur pada skala yang berbeda.

2.3 Ektraksi Fitur

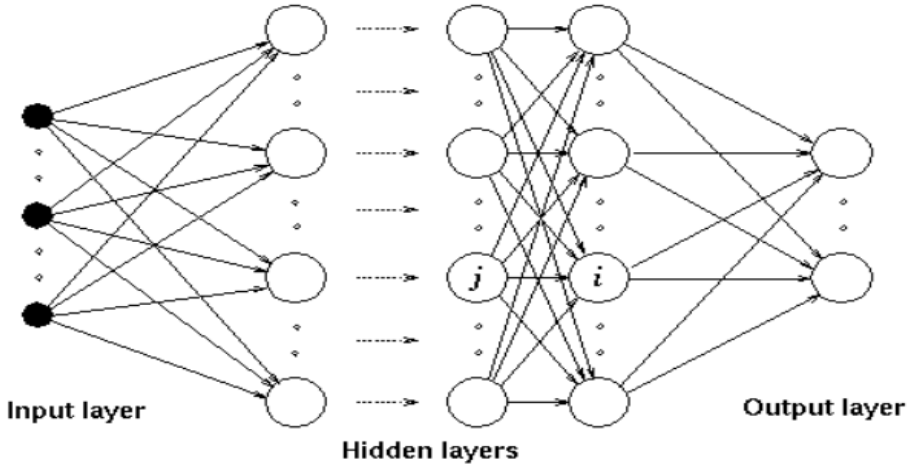
Pada penelitian ini dipergunakan pendekatan TF-IDF (Term Frequency -Inverse Document Frequency) yang merupakan salah satu metode ektraksi fitur pada data mining berbasis teks dengan melakukan perhitungan bobot kata-kata yang umum dipergunakan sehingga bisa diketahui seberapa sering suatu kata dalam hal ini kata yang berisi ujaran kasar berada pada suatu dokumen.

2.4 Klasifikasi

Klasifikasi merupakan suatu teknik yang dipergunakan untuk melakukan prediksi kata-kata yang berisi ujaran kasar. Pada Data Mining dikenal berbagai macam metode yang dipergunakan untuk melakukan prediksi, Pada penelitian ini dipergunakan 3 metode yang sering dipergunakan dalam berbagai riset, yaitu :

2.4.1 Neural Network (NN)

NN merupakan suatu metode prediksi pada data mining yang sering dipergunakan dalam berbagai jenis penelitian. NN merupakan metode yang baik dalam mentoleransi kesalahan sehingga akan menghasilkan prediksi yang tepat[13]. Bentuk arsitektur metode NN[14] dapat dilihat pada gambar 2 sebagai berikut :



Gambar 2. Arsitektur Neural Network

Pada gambar jaringan arsitektur NN diatas terdapat 3 bagian penting yaitu:

1. Input Layer
   Merupakan lapisan pertama pada NN yang berfungsi untuk memasukkan data fitur pada jaringan NN untuk diproses tahapan berikutnya.
2. Hidden Layer
   Merupakan lapisan tersembunyi pada arsitektur NN yang berfungsi untuk memproses setiap inputan dari input layer.
3. Output Layer
   Adalah lapisan terakhir pada arsitektur NN yang merupakan lapisan yang dipergunakan untuk menampilkan hasil prediksi yang telah dilakukan.

ISSN: 2685-4902 (online)

2.4.2. Decision Tree (DT)
DT merupakan suatu metode pada data mining yang juga sering dipergunakan dalam berbagai riset klasifikasi. Metode DT merupakan teknik prediksi dengan merubah bentuk data tabel menjadi model tree yang akan menghasilkan beberapa rule[15].

2.4.3. K-Nearest Neigbour (KNN)
KNN merupakan metode data mining yang cukup popular dan sering dipergunakan oleh beberapa peneliti. KNN mempunyai algoritma yang lebih sederhana[16], Metode ini relative mudah dipahami karena melakukan klasifikasi berdasar obyek terdekat dengan obyek lain disebelahnya.

HASIL DAN PEMBAHASAN

Penelitian mengenai deteksi ujaran kasar pada sosial media ini merupakan penelitian experimental dengan menggunakan Orange yang merupakan aplikasi yang dipergunakan pada pemrosesan machine learning untuk memvisualisasikan serta menganalisa dataset. Tahapan pemrosesan pada aplikasi orange dapat dijelaskan sebagai berikut :

3.1 Pengaturan Dataset

Penelitian ini menggunakan dataset ujaran kasar pada twitter dengan pengaturan fitur seperti pada tabel 1 sebagai berikut :

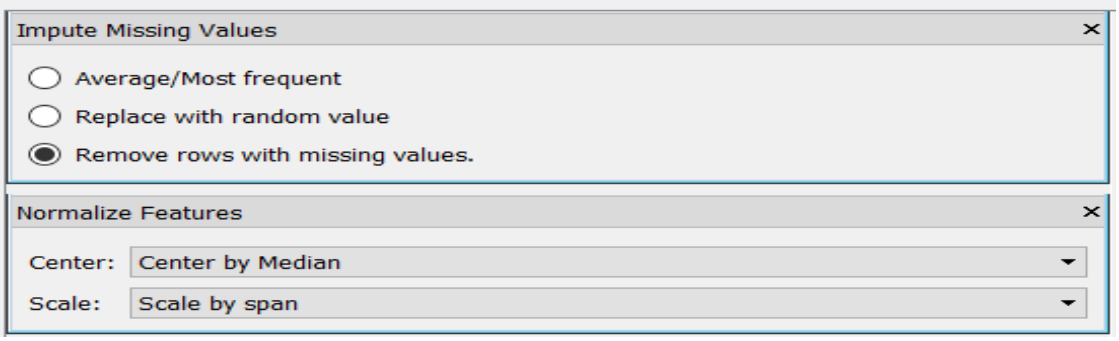
Tabel 1 Setting Dataset

Nama	Tipe	Jenis	Nilai
HS	Categorical	Feature	0,1
Abusive	Categorical	Target	0,1
HS_Ind	Categorical	Feature	0,1
HS_Grp	Categorical	Feature	0,1
HS_Rel	Categorical	Feature	0,1
HS_Rac	Categorical	Feature	0,1
HS_Phy	Categorical	Feature	0,1
HS_Gen	Categorical	Feature	0,1
HS_Oth	Categorical	Feature	0,1
HS_Wek	Categorical	Feature	0,1
HS_Mod	Categorical	Feature	0,1
HS_Str	Categorical	Feature	0,1
Tweet	Categorical	Feature	0,1

Pada tabel 1 diatas bisa terlihat bahwa Abusive merupakan label yang disetting menjadi Target sedangkan data lain merupakan fitur dengan sejumlah 12 fitur yang membentuk dataset ini.

3.2. Preprocessing Data

Preprocessing merupakan tahapan yang sangat penting pada suatu penelitian. Dengan Teknik preprocessing yang baik makaa akan dieproleh data yang bersih dan berkualitas untuk diproses pada tahap berikutnya. Pada tahapan ini dilakukan proses preprocessing seperti ditunjukkan pada gambar 3 sebagai berikut :



Gambar 3 Tahapan Data Preprocessing

3.3. Extraksi Fitur

Pada penelitian ini digunakan pendekatan TF-IDF (Term Frequency-Inverse Document Frequency) yang merupakan salah satu metode ekstraksi ciri pada data mining berbasis teks dengan menghitung bobot kata-kata yang umum digunakan sehingga dapat diketahui seberapa sering suatu kata dalam hal ini adalah kata yang digunakan. mengandung kata-kata kasar dalam sebuah dokumen.

3.4 Klasifikasi

Tahapan klasifikasi untuk melakukan deteksi pada kata-kata yang memiliki ujaran kasar dilakukan dengan menggunakan metode Neural Network. Metode NN dipergunakan karena mempunyai beberapa kelebihan yaitu pemrosesan pararel,pembelajaran mandiri,toleransi kesalahan yang tinggi serta kemampuan nonlinier yang sangat baik[17]. Menggunakan aplikasi orange kinerja dari metode NN akan dibandingkan dengan 2 metode lain yaitu KNN dan Decision Tree.

3.5. Evaluasi

Hasil pengukuran kinerja menggunakan Akurasi,Presisi dan AUC (Area Under Curve) yang hasilnya bisa terlihat pada tabel 2 sebagai berikut :

Tabel 2 Perbandingan Kinerja Metode Machine Learning

Metode	Akurasi	Presisi	AUC
Neural Network	**73 %**	**72.5%**	**73.5%**
Decision Tree	72.7%	72.4%	73.5%
KNN	68.5%	68.2%	69.8%

Tabel 2 diatas menunjukkan bahwa metode Neural Network(NN) memiliki kinerja yang paling baik dibuktikan dengan adanya nilai Akurasi,Presisi maupun AUC melebihi pendekatan lainnya. Namun jika dilihat metode Decesion Tree ternyata juga memiliki kinerja yang baik pada sisi AUC dengan nilai sama dengan NN yaitu 73.5%. AUC merupakan pendekatan dalam mengukur keakuratan suatu model pada prediksi, DecesionTree mempunyai nilai kinerja AUC yang baik karena proses perhitungan yang lebih efisien dimana sampled data diuji hanya berdasar kriteria atau kelas tertentu saja.

ISSN: 2685-4902 (online)

3.6. Confusion Matrix (CM)

CM merupakan suatu bentuk tabel yang dipergunakan dalam mengukur kinerja atau tingkat kebenaran pada klasifikasi. CM terdiri dari perhitungan beberapa nilai yang divisualisasikan pada tabel 3 dibawah ini :

Tabel 3 Confusion Matrix

TN	FP
FN	TP

Untuk mengukur kinerja model pada umumnya menggunakan akurasi. Pendekatan ini dapat dihitung dengan mempergunakan rumus perhitungan sebagai berikut :

$$\frac{TP + TN}{TP + TN + FP + FN} \text{ x } 100\%$$

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

Perhitungan CM pada masing-masing metode dapat dihitung sebagai berikut:

Tabel 4 CM Neural Network

	0	1	∑
0	6524	1602	8126
1	1945	3098	5043
∑	8469	4700	13169

Akurasi = $\frac{3098+6524}{3098+6524+1602+1945}$ x 100%

Akurasi = 73.06 %

Tabel 5 CM Decission Tree

	0	1	∑
0	6535	1591	8126
1	2003	3040	5043
∑	8538	4631	13169

Akurasi = $\frac{3040+6535}{3040+6535+1591+2003}$ x 100%

Akurasi = 72.70 %

Tabel 6 CM KNN

	0	1	∑
0	6169	1957	8126

1	2194	2849	5043
∑	8363	4806	13169

Akurasi = $\frac{2849+6169}{2849+6169+1957+2194}$ x 100%

Akurasi = 68.47 %

Pada perhitungan CM pada tabel 4, 5 dan 6 diatas label angka 0 menunjukkan kata yang bukan merupakan ujaran kasar sedangkan label angka 1 menunjukkan kata yang berisi ujaran kasar.

KESIMPULAN

Penelitian mengenai deteksi ujaran kasar pada sosial media terutama Twitter merupakan riset yang sangat menarik karena pertumbuhan pengguna sosial media yang semakin meningkat dari tahun ke tahun serta permasalahan ujaran kasar berpotensi menyebabkan pertikaian dan pelanggaran hukum. Penggunaan metode yang tepat dalam melakukan deteksi ujaran kasar akan sangat membantu bagi interaksi sosial media pada masyarakat. Pengolahan data ujaran kasar berbasis teks pada Twitter dalam penelitian ini menggunakan metode preprocessing yang tepat serta menggunakan Neural Network sebagai salah satu metode yang unggul pada Machine Learning. Hasil penelitian membuktikan bahwa NN mempunyai tingkat akurasi tertinggi sebesar 73 % lebih baik dibandingkan metode lain seperi Decesion Tree maupun KNN untuk mendeteksi ujaran kasar pada sosial media.

DAFTAR PUSTAKA

[1] M. A. Harahap and S. Adeni, "Tren penggunaan media sosial selama pandemi di indonesia," *J. Prof. FIS UNIVED*, vol. 7, no. 2, pp. 13–23, 2020.

[2] C. Juditha, "Fenomena Trending Topic Di Twitter: Analisis Wacana Twit #Savehajilulung Trending Topic Phenomenon on Twitter: Discourse Analysis of Tweet #Savehajilulung," *Penelit. Komun. dan Pembang.*, vol. 16, no. 2, pp. 138–154, 2015.

[3] Sasongko, V. A. A. Artanti, N. U. Putri, J. Hendrawan, and S. D. Sari, "Ujaran Kebencian di Media Sosial dalam Perspektif Cyberlaw di Indonesia," *Proceeding Conf. Law Soc. Stud.*, pp. 1–12, 2021, [Online]. Available: http://prosiding.unipma.ac.id/index.php/COLaS

[4] M. Hakiem, M. A. Fauzi, and Indriati, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2443–2451, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4682

[5] Willianto, I. A. Musdar, Junaedy, and H. Angriani, "Implementasi Teori Naive Bayes Dalam Klasifikasi Ujaran Kebencian," *J. Inform. Univ. Pamulang*, vol. 6, no. 4, pp. 666–671, 2021.

[6] Oryza Habibie Rahman, Gunawan Abdillah, and Agus Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 17–23, 2021, doi: 10.29207/resti.v5i1.2700.

[7] A. N. Ulfah and M. K. Anam, "Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[8] J. Media and I. Budidarma, "Feature Expansion Using Word2vec for Hate Speech Detection on Indonesian Twitter with Classification Using SVM and Random," vol. 6, no. April, pp. 979–988, 2022, doi: 10.30865/mib.v6i2.3855.

ISSN: 2685-4902 (online)

[9] F. Ihsan, I. Iskandar, N. S. Harahap, and S. Agustian, "Decision tree algorithm for multi-label hate speech and abusive language detection in Indonesian Twitter," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 4, pp. 199–204, 2021, doi: 10.14710/jtsiskom.2021.13907.
[10] S. Malmasi and M. Zampieri, "Detection of Hate Speech in Indonesian Language on Twitter Using Machine Learning Algorithm Febby," *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, vol. 2017-Septe, pp. 467–472, 2017, doi: 10.26615/978-954-452-049-6-062.
[11] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif*. Alfabeta, CV, 2017.
[12] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *Proc. Third Work. Abus. Lang.*, pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.
[13] H. Putra and N. Ulfa Walmi, "Penerapan Prediksi Produksi Padi Menggunakan Artificial Neural Network Algoritma Backpropagation," *J. Nas. Teknol. dan Sist. Inf.*, vol. 6, no. 2, pp. 100–107, 2020, doi: 10.25077/teknosi.v6i2.2020.100-107.
[14] B. S. Santoso, J. P. Tanjung, U. P. Indonesia, B. Gandum, and A. N. Network, "Classification of Wheat Seeds Using Neural Network Backpropagation," *JITE (Journal Informatics Telecommun. Eng. Available*, vol. 4, no. January, pp. 188–197, 2021.
[15] S. Bahri, A. Lubis, U. Pembangunan, and P. Budi, "Metode Klasifikasi Decision Tree Untuk Memprediksi Juara English Premier League," *Sintaksis*, vol. 2, no. 04, pp. 63–70, 2020.
[16] L. Farokhah, "Implementasi K-Nearest Neighbor untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna RGB," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, p. 1129, 2020, doi: 10.25126/jtiik.2020722608.
[17] Y. Lai, "Application of machine learning algorithm based on neural network technology," *J. Phys. Conf. Ser.*, vol. 2066, no. 1, 2021, doi: 10.1088/1742-6596/2066/1/012041.