p-ISSN:2527-5321
e-ISSN:2527-5941

DETEKSI CYBERBULLYING TWEET MENGGUNAKAN MACHINE LEARNING

Maulina Safitri[1], Najwa Alhaura Zafira[2], Adinda Dyahrestu Putri[3], Sabrina Putri Pamungkas[4], Fatimah Azahrah[5], Ajeng Lestari[6], Diajeng Kusdiyah[7], Ni Wayan Parwati[8]

Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer
Universitas Indraprasta PGRI
Jalan Raya Tengah No 80, Kelurahan Gedong, Pasar Rebo, Jakarta Timur

maulinasafitri62@gmail.com [1], najwaalhaurazafira23@gmail.com [2], adindadyahrestuputri04@gmail.com [3], sabrinaputri2002.sp@gmail.com [4], fatimah.rraa@gmail.com [5], ajengles03@gmail.com [6], diajengkusdiyah@gmail.com [7], wayan.parwati@gmail.com [8]

ABSTRAK

Menurut The United Nations Children's Fund (UNICEF) cyberbullying merupakan penindasan ataupun perundungan dengan memakai teknologi digital. Perihal ini bisa terjalin di media sosial, platform pengiriman pesan, platform permainan, ataupun telepon seluler. Aksi ini dicoba secara kesekian yang diperuntukan buat menakut- nakuti, membuat marah, ataupun mempermalukan orang- orang yang jadi sasaran. Cyberbullying merupakan aksi yang sangat merugikan serta dapat mempunyai akibat emosional serta psikologis yang sungguh- sungguh pada korban. Penting untuk mengambil tindakan untuk melindungi diri sendiri dan melaporkan tindakan tersebut kepada pihak berwenang atau penyedia platform media sosial. Perihal ini butuh dicoba oleh korban ataupun siapa juga yang melihat aksi tersebut. Penelitian ini mengklasifikasikan data kedalam kelas cyberbullying dan not cyberbullying menggunakan metode machine learning, yaitu BiLSTM (Bidirectional Long Short-Term Memory), BERT (Bidirectional Encoder Representations from Transformers), dan RF (Random Forest), untuk mengetahui cyberbullying atau bukan. Ditemukan bahwa BERT mencapai akurasi sebesar 94%, serta f1 score 89% sehingga lebih unggul dalam menanggulangi ketidakseimbangan data. Hasilnya menampilkan kalau akumulasi penambahan kata kunci memakai TF-IDF dan borda rangking hasil ekstraksi kata dapat meningkatkan akurasi sampai 80%. Sentimen analisis memakai Majority Voting, K-Means Clustering, dan BERT menunjukkan hasil akurasi 83%, dengan label sentimen positif, negatif, sangat positif, dan sangat negatif.
Kata Kunci: *analisis sentimen, cyberbullying, machine learning, majority voting, topic modelling, sentiment analysis*

ABSTRACT

According to the United Nations Children's Fund (UNICEF), cyberbullying refers to bullying or harassment using digital technology. This can occur on social media, messaging platforms, gaming platforms, or mobile phones. These actions are repeatedly carried out with the intent to intimidate, anger, or embarrass individuals who are the targets. Cyberbullying is a harmful act that can have serious emotional and psychological consequences for the victim. It is important to take action to protect oneself and report such behavior to authorities or the social media platform provider. This should be done by the victim or anyone who witnesses the incident. This study classifies data into cyberbullying and non-cyberbullying categories using machine learning methods, including (Bidirectional Long Short-Term Memory, Bidirectional Encoder Representations from Transformers, and Random Forest, to detect whether it is cyberbullying or not. The results show that BERT achieved an accuracy of 94% and an F1 score of 89%, making it more effective in addressing data imbalance. Additionally, combining keyword enhancement using TF-IDF and ranking results from word extraction increased accuracy up to 80%. Sentiment analysis using Majority Voting, K-Means Clustering, and BERT yielded an accuracy of 83%, with sentiment labels categorized as positive, negative, very positive, and very negative.
Keywords*: analisis sentimen, cyberbullying, machine learning, majority voting, topic modelling, sentiment analysis*

PENDAHULUAN

Di era digital saat ini, media sosial telah mengubah cara orang berkomunikasi, berinteraksi, dan mengekspresikan diri secara mendasar. Platform seperti Twitter adalah ruang virtual yang menghubungkan jutaan orang dari berbagai wilayah dan demografi,

p-ISSN:2527-5321
e-ISSN:2527-5941

menjadikannya mikrokosmos dinamis dari interaksi sosial global. Indonesia merupakan negara yang sangat menarik untuk dicermati dalam konteks ini, mengingat Indonesia dianggap sebagai salah satu kekuatan digital terkemuka di kawasan Asia Tenggara.

Jumlah di Indonesia sangat mengesankan, menurut data komprehensif dari Kementerian Komunikasi dan Informatika. Dari sekitar 63 juta pengguna internet, 95% merupakan pengguna aktif media sosial. Twitter khususnya merupakan platform yang menarik banyak perhatian. Country Industry Chief Twitter Indonesia bahkan mengklaim bahwa Indonesia adalah negara dengan pertumbuhan pengguna aktif Twitter harian paling mengesankan di dunia.

Namun selain menjanjikan kemungkinan - kemungkinan positif, media sosial juga membuka wilayah yang kompleks dan berpotensi berbahaya. Penggunaan Twitter tidak lagi terbatas pada individu, namun telah meluas hingga mencakup instansi pemerintah, komunitas, dan ekosistem bisnis online. Sayangnya, tidak semua pengguna menggunakan teknologi ini dengan bijak dan bertanggung jawab.

Fenomena negatif seperti penyebaran berita palsu, penipuan, ujaran kebencian, dan yang paling parah, cyberbullying telah menjadi ancaman serius dalam ekosistem digital. Data pemerintah yang mengejutkan menunjukkan bahwa 84% remaja Indonesia berusia 12 hingga 17 tahun pernah menjadi korban perundungan, dan sebagian besar terjadi secara online. Penindasan di media sosial sering kali terjadi melalui tweet yang berisi bahasa yang menyinggung atau serangan berdasarkan ras, etnis, atau agama.

Tweet sendiri merupakan salah satu media sosial terfavorit di dunia dari We Are Social dan Hootsuite terbaru, terdapat sebanyak 5,35 miliar pengguna internet dan 5,05 miliar pengguna media sosial di seluruh dunia per Februari 2024, dan We Are Social tahun 2023, pengguna media sosial dapat mencapai 167 juta jiwa dari jumlah populasi sebesar 276.4 juta jiwa. Pada tahun pemilu ini, Indonesia telah mengirimkan 42,1 juta tweet terkait pemilu sejauh ini, meningkat 36 persen dibandingkan periode yang sama tahun lalu. Secara global, Twitter memiliki 255 juta pengguna aktif bulanan dalam angka terbaru yang dilaporkan sendiri yang terungkap pada bulan April.

UNICEF mendefinisikan cyberbullying sebagai bentuk penindasan yang dilakukan menggunakan teknologi digital dan dapat terjadi melalui telepon, platform pesan, media sosial, dan platform game. Tindakan ini diulangi dengan tujuan untuk menakut-nakuti, membuat marah, atau mempermalukan korban dan dapat berdampak serius pada kesehatan emosional dan psikologis mereka.

Mengingat kompleksitas masalah ini, penelitian kami mengusulkan pendekatan inovatif yang memanfaatkan pembelajaran mesin sebagai solusi strategis untuk mendeteksi dan memitigasi penindasan maya. Dengan menggunakan algoritma canggih seperti regresi linier dan hutan acak, kami bertujuan untuk mengembangkan model deteksi otomatis yang dapat mengidentifikasi konten berbahaya secara akurat dan efektif.

Tujuan khusus dari penelitian ini adalah: Mengembangkan model deteksi otomatis untuk mengidentifikasi tweet yang mengandung cyberbullying, menentukan algoritma pembelajaran mesin yang paling akurat untuk mendeteksi konten negatif, menggunakan metrik presisi, presisi, dan recall, tantangan dan keterbatasan sistem deteksi analitis cyberbullying. Melalui pendekatan ilmiah dan teknologi ini, kami ingin memerangi penindasan maya, melindungi pengguna media sosial, dan memberikan kontribusi nyata untuk membangun ruang digital yang lebih aman, inklusif, dan bermartabat.

METODE PENELITIAN

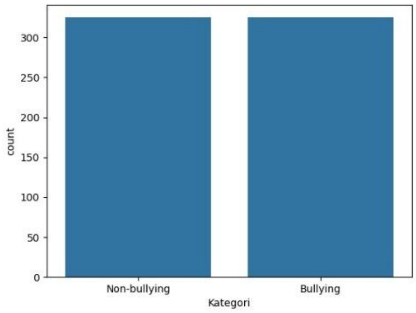
Penelitian ini menggunakan machine learning untuk membandingkan algoritma klasifikasi teks untuk mengidentifikasi cyberbullying berdasarkan postingan Twitter pengguna. Tahapan penelitian ini digambarkan pada Gambar 1.1

Data Collection

Dataset yang digunakan adalah data sekunder yang diambil dari dataset Cyberbullying

p-ISSN:2527-5321
e-ISSN:2527-5941

Kaggle (link). Dataset ini terdiri dari lima atribut, yaitu teks, label, usia, gender, dan kategori usia, dan diekstraksi menggunakan API Tweet. Tabel 1 menunjukkan lima atribut dari dataset yang digunakan, dengan label 0 menunjukkan negatif dan label 1 menunjukkan positif.



Gambar 1.1 Data Collection

Preprocessing

Preprocessing membuat data yang akan diproses lebih terstruktur dan pemodelan menjadi lebih mudah. Lima langkah preprocessing yang digunakan dalam penelitian ini adalah pembersihan teks, penyusunan case, tokenisasi, filtering, dan stemming. Tujuan dari proses pembersihan teks adalah untuk menghilangkan angka, pemisah kata, seperti koma, titik, dan tanda baca lainnya. Pengolahan case adalah langkah preprocessing yang bertujuan untuk mengubah semua teks dalam dokumen menjadi huruf kecil. Tokenisasi menghapus string yang dimasukkan. Untuk memisahkan kalimat menjadi kelompok kata, spasi dan tanda baca dihilangkan. Stopword removal menghilangkan kata-kata yang tidak relevan atau tidak penting. Stemming adalah proses mencari kata dasar (stem word) yang dihasilkan dari proses penghapusan stopword. Setelah tahap preprocessing, metode TF-IDF digunakan untuk menghitung bobot setiap kata dalam setiap dokumen.

TF-IDF mengubah kata pada setiap dokumen menjadi angka, yang kemudian disusun menjadi sebuah matriks. Nilai kemunculan kata (Term Frequency) pada setiap dokumen dihitung.

Feature Extraction

Dalam penelitian ini, perhitungan vektor yang dilakukan menggunakan metode CountVectorizer adalah hasil dari fitur kelas perhitungan numerik dan metode ekstraksi fitur teks yang umum digunakan. CountVectorizer mengubah teks menjadi matriks frekuensi kata, dan kemudian fungsi matriks fit_transform digunakan untuk menghitung jumlah kali setiap kata muncul. Tujuannya adalah untuk mengumpulkan kata perkata dari setiap kalimat dan membuat vektor fitur yang terdiri dari banyak kata.

Splitting Data

Setelah tahap preprocessing selesai, kumpulan data dibagi menjadi dua bagian: data uji dan data pelatihan. Ini dilakukan dengan menggunakan pemisahan data 80:20. Data pelatihan adalah jumlah yang digunakan untuk melakukan penelitian, dan data penguji adalah jumlah yang belum pernah digunakan dalam penelitian tetapi berguna untuk mengevaluasi keberhasilan atau kegagalan penelitian.

Machine Learning

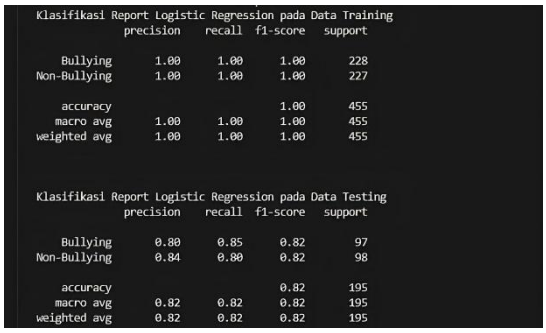
Salah satu komponen kecerdasan buatan adalah pembelajaran mesin, yang dimaksudkan untuk memungkinkan mesin melakukan tugasnya dengan cara yang cerdas. Algoritma pembelajaran mesin perlu terhubung ke database untuk melakukan pengindeksan data. Tujuan utama model pembelajaran mesin adalah menggunakan komputer untuk membuat prediksi, tetapi itu juga terkait dengan statistik komputer. Random forest dan algoritma pembelajaran mesin linier regression.

HASIL DAN PEMBAHASAN

Dataset publik yang dikumpulkan dari website Kaggle terdiri dari 650 tweet yang berkaitan dengan cyberbullying yang ditemukan di jejaring sosial Twitter. Data ini penting bagi penelitian ini. Preprocessing dilakukan sebelum menganalisis data. Ini mencakup menghilangkan data null atau kosong, menghilangkan data duplikat, mengubah font semua teks dalam dokumen menjadi font yang konsisten, mengubah semua kalimat dalam dokumen menjadi unit kata, menghilangkan kata, simbol, atau URL yang tidak berguna, dan mengubah data untuk memenuhi kebutuhan algoritma. Setelah melakukan preprocessing data dan ekstraksi fitur Semua kata dalam teks akan diperiksa satu per satu untuk memastikan apakah ada

p-ISSN:2527-5321
e-ISSN:2527-5941

kata-kata yang ditulis dengan cara yang tidak biasa. Kata-kata akan diganti dengan tulisan yang sebenarnya jika ditemukan. Membuat kamus khusus yang disesuaikan dapat digunakan untuk normalisasi teks.



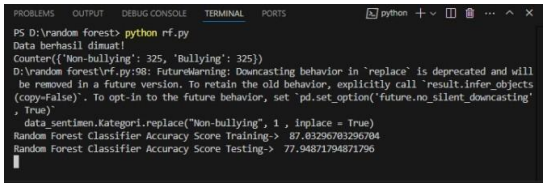
Klasifikasi Report Logistic Regression pada Data Training

	precision	recall	f1-score	support
Bullying	1.00	1.00	1.00	228
Non-Bullying	1.00	1.00	1.00	227
accuracy			1.00	455
macro avg	1.00	1.00	1.00	455
weighted avg	1.00	1.00	1.00	455

Klasifikasi Report Logistic Regression pada Data Testing

	precision	recall	f1-score	support
Bullying	0.80	0.85	0.82	97
Non-Bullying	0.84	0.80	0.82	98
accuracy			0.82	195
macro avg	0.82	0.82	0.82	195
weighted avg	0.82	0.82	0.82	195

Gambar 2.1 (Logistic Regression)



Gambar 2.2 (Random Forest)

Dalam program ini, algoritma regresi logistik dan random forest yang memiliki tingkat akurasi tertinggi adalah regresi logistik, dengan tingkat akurasi sebesar 87%. Ini dibuktikan dengan nilai tes akurasi sebesar 83.08 persen, precision, recall, dan FI-Score bullying dan non-bullying sebesar 1.00:0.99, 0.99:1.00, dan 1.00:1.00, dengan nilai pembagian data 80:20. Data yang telah melewati beberapa langkah akan kemudian didistribusikan.

SIMPULAN DAN SARAN

Hasilnya menunjukkan bahwa melalui klasifikasi dataset, penelitian telah berhasil mengidentifikasi korelasi signifikan antara cyberbullying pada Twitter. Menurut hasil evaluasi machine learning algoritma, regresi logistik lebih baik daripada hutan random dalam mengklasifikasi data terkait ini. Ini ditunjukkan oleh nilai tes ketepatan 83,08 persen; nilai tes ketepatan, recall, dan FI-score bullying dan non-bullying masing-masing 1,00:0,99, 1,00:0,99, dan 1,00:1,00.

Karena penelitian saat ini terbatas pada model klasifikasi cuitan yang menunjukkan cyberbullying pada media sosial, penulis berharap dapat melanjutkan penelitian dengan menambahkan data pelatihan dan antarmuka pengujian model.

UCAPAN TERIMA KASIH

Terima kasih kepada semua pihak yang telah berkontribusi dalam proses pembuatan program DETEKSI CYBERBULLYING TWEET MENGGUNAKAN MACHINE LEARNING ini. Hasil penelitian ini tidak akan mungkin tercapai tanpa bimbingan dari dosen pembimbing yaitu Ibu Ni Wayan Parwati, S.Kom., dukungan rekan-rekan magang, serta bantuan dari pihak-pihak terkait. Setiap data, informasi, dan masukan yang diberikan sangat berarti dalam menyempurnakan hasil penelitian ini. Semoga hasil penelitian ini dapat memberikan manfaat dan menjadi kontribusi yang berharga bagi pengembangan di bidang ini. Terima kasih atas segala dukungan dan kerjasamanya.

DAFTAR PUSTAKA

Fauzan, B.L., Agustin, T., dan Mahmudah, A.M.H. (2024). Metode Regresi Logistik Multinomial digunakan untuk Memprediksi Klasifikasi Kecelakaan Lalu Lintas di Kota Surakarta. Journal of Sustainable Civil Building Management and Engineering, Vol. 1(4),1-9. https://journal.pubmedia.id/index.php/civilengineering

Hootsuite (We are Social): Data Digital Indonesia 2024. (n.d.). Dosen, Praktisi, Konsultan, Pembicara/Fasilitator Digital Marketing, Internet Marketing, SEO, Technopreneur Dan Bisnis Digital. Retrieved October 16, 2024, from https://andi.link/hootsuite-we-are-social-data-digital-indonesia-2024/#google_vignette

Hootsuite (We are Social): Indonesian Digital Report 2023. (n.d.). Dosen, Praktisi, Konsultan, Pembicara/Fasilitator Digital Marketing, Internet Marketing, SEO, Technopreneur Dan Bisnis Digital. Retrieved October 16, 2024, from https://andi.link/hootsuite-we-are-social-indonesian-digital-report-2023/

Hu, J. dan Szymczak, S. ulasan analisis jangka panjang dengan hutan random Institut Biometri dan Statistik Medis

p-ISSN:2527-5321
e-ISSN:2527-5941

Universitas Lübeck. Sumber: silke.szymczak@uni-luebeck.de

Mahmud AF dan Wirawan S. (2024). Metode klasifikasi machine learning digunakan untuk mendeteksi web phishing. Jurnal Sistem Informasi, 13(4), 1368-1380, dapat ditemukan di http://sistemasi.ftik.unisi.ac.id.

Rahayu, K., Fitria, V., Septhya, D., Rahmaddeni, & Efrizoni, L. (2023). *Text* classification for detecting depression and anxiety among Twitter users based on machine learning. Program Studi Teknik Informatika, STMIK Amik Riau, Pekanbaru, Riau. Received June 4, 2023; Revised July 14, 2023; Accepted August 20, 2023.

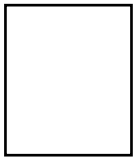
Sari. A. K., Irsyad, A., Aini, D N., Islamiyah, & Ginting , S. E. (n.d.). Analisis Sentimen Twitter Menggunakan Machine Learning untuk Identifikasi Konten Negatif. Program Studi Sistem Informasi, Universitas Mulawarman

Tech in Asia - Connecting Asia's startup ecosystem. (n.d.). Tech in Asia. Retrieved October 16, 2024, from https://www.techinasia.com/twitter-close-20-million-active-users-indonesia

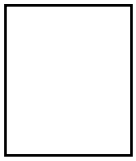
Wijanto, M. C. (2015). *Sistem pendeteksi pengirim tweet dengan metode klasifikasi* chaeNIDN: 0323098401Naive Bayes. Jurnal Teknik Informatika dan Sistem Informasi, Vol. 1(2), 172. https://doi.org/ISSN2443-2229.

zefanyasosiawan@student.telkomuniversity.ac.id. (2024, July 14). D3 Teknologi Telekomunikasi. D3 Teknologi Telekomunikasi. https://dte.telkomuniversity.ac.id/cyber-bullying/

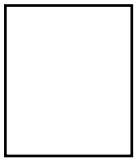
Biografi Penulis



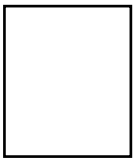
Fatimah Azahrah
202143501356



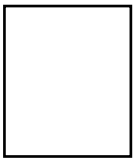
Sabrina Putri Pamungkas
202143501361



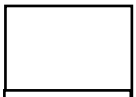
Diajeng Kusdiyah -
202143501335



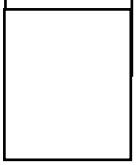
Ajeng Lestari
202143501337



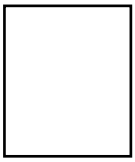
Maulina Safitri
202143501374



Adinda Dyahrestu Putri
202143502663



Najwa Alhaura Zafira
202143501385



Ni Wayan Parwati Septiani, M.M., M.Kom
NIDN: 0323098401.