Hate Speech and Emotion Detection on Twitter Using LSTM Model

Deteksi Ujaran Kebencian dan Emosi di Twitter Menggunakan Model LSTM

Nanda Yunania[1], Yulian Findawati[2]*

[1,2] Program Studi Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

Abstract. This research aims to develop a classification model to detect hate speech and emotions on the Twitter platform used the Long-Short Term Memory (LSTM) method. With the increasing volume of data on social media, especially Twitter, automatic identification of negative content is crucial for maintaining a healthy digital ecosystem. The dataset used in this study consists of tweets labeled for hate speech and various emotion categories. The preprocessing process is carried out to clean and prepare the data, including steps such as punctuation removal, tokenization, and text normalization. After preprocessing, the dataset is split into training and testing data with a ratio of 60:40 to ensure accurate model evaluation. The experimental results show that the LSTM model achieves an accuracy of 89% in hate speech classification and 71% in emotion classification. These results demonstrate the potential of the LSTM method in text analysis tasks and can serve as a basis for developing automatic detection systems on social media platforms.

Keywords : Hate Speech, LSTM, Twitter

Abstrak. Penelitian ini bertujuan untuk mengembangkan model klasifikasi untuk mendeteksi hate speech dan emosi dalam platform Twitter dengan metode Long-Short Term Memory (LSTM). Dengan semakin meningkatnya volume data di media sosial, khususnya Twitter, identifikasi otomatis terhadap konten bermuatan negatif menjadi sangat penting untuk menjaga ekosistem digital yang sehat. Dataset yang digunakan dalam penelitian ini terdiri dari tweet yang telah diberi label untuk hate speech dan berbagai kategori emosi. Proses preprocessing dilakukan untuk membersihkan dan menyiapkan data, termasuk langkah-langkah seperti penghapusan tanda baca, tokenisasi, dan normalisasi teks. Setelah preprocessing, dataset dibagi menjadi data latih dan data uji dengan proporsi 60:40 untuk memastikan evaluasi model yang akurat. Hasil eksperimen menunjukkan bahwa model LSTM mencapai akurasi sebesar 89% dalam klasifikasi hate speech dan 71% dalam klasifikasi emosi. Hasil ini menunjukkan potensi metode LSTM dalam tugas analisis teks dan dapat digunakan sebagai dasar untuk pengembangan sistem deteksi otomatis di platform media sosial.

Kata Kunci : Ujaran Kebencian, LSTM, Twitter

OPEN ACCESS

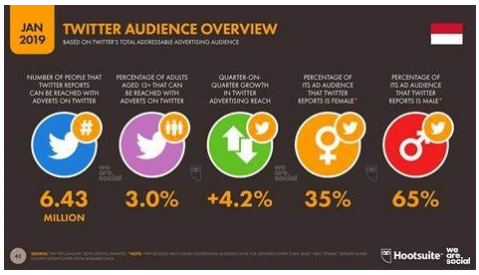
ISSN 2503 3492 (online)

**Correspondence:*
Yulian Findawati
yulianfindawati@umsida.ac.id

Citation:
Nanda Yunania, Yulian Findawati (2023) Hate Speech and Emotion Detection on Twitter Using LSTM Model. Journal of Information and Computer Technology Education. 7i1. doi:10.21070/jicte.v7i1.1645

PENDAHULUAN

Banyaknya ragam aplikasi sosial media, banyak warga Indonesia yang menggunakan media Twitter sebagai media komunikasi dan mencari informasi. Platform Twitter adalah situs web dan aplikasi seluler yang digunakan untuk mengakses dan berinteraksi dengan layanan media sosial Twitter. Twitter juga dapat diakses melalui berbagai perangkat seperti komputer, ponsel pintar, dan tablet. Pengguna dapat membuat akun di platform ini, mengirim tweet, mengikuti akun lain, dan terlibat dalam berbagai aktivitas seperti berbagi informasi, berdiskusi, dan berpartisipasi dalam percakapan yang beragam. Selain itu, Twitter juga memiliki berbagai fitur tambahan, seperti fitur live streaming dan iklan, yang memungkinkan pengguna untuk berinteraksi dengan berbagai konten dan iklan (Aldi et al., 2018a).



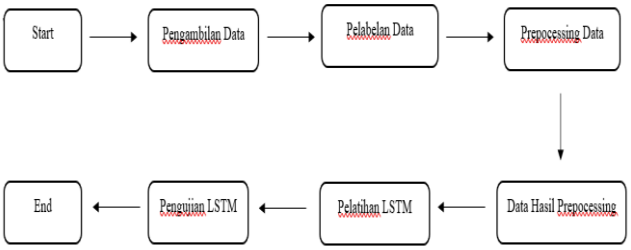
Gambar 1. Perkembangan Media Sosial di Indonesia

Ujaran kebencian (Hate Speech) adalah jenis komunikasi atau ekspresi yang memiliki niat atau tujuan untuk menyakiti, merendahkan, atau merendahkan kelompok atau individu berdasarkan atribut seperti ras, etnisitas, agama, gender, orientasi seksual, atau faktor lainnya (Kholifatullah & Prihanto, 2023). Ada berbagai macam jenis-jenis Hate Speech pada platform Twitter yang didasari oleh beberapa motif-motif tertentu, antara lain motif suku, agama, ras, dan antar golongan atau bisa disebut SARA (Murni et al., 2023).

Dan juga ada Ujaran emosi, juga dikenal sebagai teks emosi, merujuk pada teks yang mengandung ekspresi emosional, seperti kemarahan, kegembiraan, kesedihan, atau kecemasan. Analisis ujaran emosi telah digunakan dalam berbagai konteks, termasuk analisis media sosial, survei pendapat, dan pemantauan citra merek, untuk memahami persepsi publik terhadap berbagai topik atau merek (Mardia et al., 2023)

METODE

Metodologi penelitian merupakan suatu gambaran yang didalamnya terdapat proses yang digagas dalam penelitian ini. Gambaran itu diperlukan untuk contoh guna memiliki hasil yang memenuhi target dari penelitian ini.(Hartono, 2017) Pada penelitian ini, penulis mengimplementasikan metode Long Short Term Memory (LSTM) pada studi kasus klasifikasi hate speech dan emosi pada platfrom Twitter. Berikut langkah-langkah yang dilakukan:



a. Data Set merupakan himpunan data, terdiri dari kumpulan objek dan atributnya. Istilah umum yang biasanya digunakan untuk objek meliputi rekam, titik, vektor, pola, peristiwa, observasi, kasus, sampel, instansi, dan entitas.(Radliya, 2015)
b. Prepocessing merupakan hal penting karena data pada dunia nyata kadang tidak selalu lengkap (terkadang ada kurang nilai atribut, kurang atribut tertentu yang menarik, ataupun hanya mengandung data agregat), kotor atau noise (mengandung kesalahan/outlier), dan tidak konsisten (berisi perbedaan dalam kode atau nama). Yang dibagi menjadi beberapa proses antara lain : Case Folding, Simbol & Whitespace Removal, Normalisasi Kata, Tokenizing, Lemmatization, Stopword Removal (Prasetyo, 2012).
c. Split Data adalah proses membagi dataset menjadi subset-subset yang berbeda, yang biasanya digunakan untuk pelatihan, validasi, dan pengujian model dalam pembelajaran mesin dan analisis data (Kadir, 2021).
d. Pemodelan LSTM dikembangkan untuk mengatasi masalah hilangnya informasi jangka panjang dalam jaringan saraf rekurensi konvensional. Dengan adanya "gerbang" khusus dalam struktur LSTM, model ini dapat mempelajari dan mengingat informasi jangka panjang dengan lebih efektif. Antara lain : Forget Gate, Input Gate, Output Gate (Aldi et al., 2018b).

HASIL DAN PEMBAHASAN

A. Data Set

Twitter memiliki beberapa unsur yang ada di dalamnya, antrara lain : URL, tanggal, Tweet, ID, Username, Like, Quotes, Komentar dan lainnya. Dari unsur tersebut berhasil mengumpulkan 2473 tweet. Data tersebut kemudian diolah ke tahapan selanjutnya. (Rafael & Adikara, 2023)

Tabel 1. Data Hasil Crawling

NO	Tweet
1	Met Pagi manado kasih kendor’
2	Kok sewot dgn Pidato Sambutan
…	Bpk Joko Widodo ...
2473	....
	Kasihan Ibu ini jadi korban akibat dicuci otak...

B. Pre-Processing

Preprocessing diperlukan karena data dunia nyata umumnya tidak lengkap (kurang nilai atribut, kurang atribut tertentu yang menarik, atau hanya berisi data agregat), kotor / noise (mengandung kesalahan atau outlier), dan tidak konsisten (mengandung perbedaan dalam kode atau nama). (Henderi & Wanda, 2017)

Tabel 2. Proses Preprocessing Data

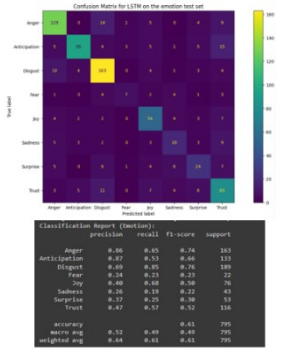
Proses	Tweet
Case Folding	met pagi manado kasih kendor’
Simbol & Whitespeace Removal	met pagi manado kasih kendor
Normalisasi Kata	selamat pagi manado
Tokenizing	jangan kasih kendor
Lemmatization	selamat, pagi, manado,
Stopword Removal	jangan, kasih, kendor
	selamat pagi manado jangan kendor
	Selamat manado kendor

C. Split Data

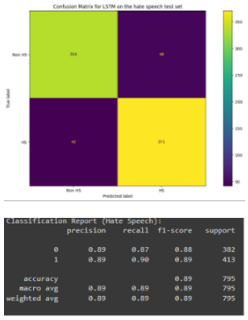
Split data adalah proses membagi dataset menjadi subset-subset yang berbeda, yang biasanya digunakan untuk pelatihan, validasi, dan pengujian model dalam pembelajaran mesin dan analisis data. Dalam konteks pembelajaran mesin, dataset umumnya dibagi menjadi set pelatihan, set validasi, dan set pengujian. Pembagian ini penting untuk menguji kinerja model pada data yang tidak terlihat sebelumnya dan memastikan bahwa model dapat digeneralisasi dengan baik pada data baru. (Wati & Pramartha, 2022)

D. Klasifikasi

Setelah melalui tahapan proses pembobotan kata, langkah berikutnya adalah memasukkan data ke dalam fase pemodelan klasifikasi. Sebelumnya, data akan dibagi menjadi dua bagian, yakni data latih dan data uji, dengan rasio 60:40, di mana 60% digunakan untuk data latih dan 40% untuk data uji dalam eksperimen ini.(Saputri et al., 2017) Data uji berperan sebagai alat evaluasi model, sementara data latih berfungsi untuk pembangunan model dan identifikasi pola. Setelah pembagian data diselesaikan, langkah selanjutnya melibatkan proses klasifikasi model. Algoritma pertama yang akan dilakukan percobaan adalah LSTM (Pradana et al., 2023)



Gambar 2. Confusion Matrix Emosi



Gambar 3. Confusion Matrix Hate Speech Set

E. Output

Output yang dihasilkan dari program ini adalah untuk mendeteksi hate speech dan emosi dari sebuah kalimat random. Ada dua jenis output dari program ini yatu berupa keterangan HS/Non HS dan emosi. (Pratama & Findawati, 2020)

Tabel 3. Output Klasifikasi Hate Speech dan Emosi

Text	*Hate Speech*	*Emosi*
Ha ha ha sigundul penguasa ancol	False	Trust
Kasihan ibu ini jadi korban akibat dicuci otak	True	Sadness

Kok sewot dengan pidato sambutan Bpk Joko Widodo	False	Anticipation
Ha ha hakadrun pada stressssssss mengenai beber	False	Trust

SIMPULAN

Dengan Output dari program ini adalah hasil klasifikasi teks tweet berbahasa Indonesia yang telah dianalisis menggunakan algoritma LSTM. Program ini mengeluarkan dua jenis klasifikasi utama:

1. Klasifikasi Hate Speech: Program akan memberikan label pada setiap tweet apakah termasuk dalam kategori ujaran kebencian atau tidak. Ini dilakukan dengan memproses teks tweet dan membandingkannya dengan pola-pola yang telah dipelajari dari dataset pelatihan.
2. Klasifikasi Emosi: Selain mendeteksi hate speech, program juga mengidentifikasi emosi yang terkandung dalam teks tweet. Emosi yang dapat dideteksi misalnya marah, sedih, bahagia, dan lain-lain. Setiap tweet akan diberi label sesuai dengan emosi yang paling dominan berdasarkan analisis fitur linguistiknya.
3. Dapat diambil kesimpulan bahwa LSTM cukup efektif dan akurat digunakan pada klasifikasi yang menghasilkan *Hate Speech* 89% dan emosi 71%

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih sebesar-besarnya kepada semua pihak yang ikut berperan dalam pembuatan jurnal ini, jurnal ini mungkin akan kurang maksimal bila tidak ada bantuan atau dorongan dari berbagai pihak yang membantu. Semoga dengan adanya jurnal ini bisa menjadi referensi baca yang berguna dengan baik bagi pembaca setelahnya.

REFERENSI

Aldi, M. W. P., Jondri, & Aditsania, A. (2018a). Analisis dan implementasi long short term memory neural network untuk prediksi harga Bitcoin. *e-Proceeding Engineering*, *5*(2), 3548–3555.

Aldi, M. W. P., Jondri, & Aditsania, A. (2018b). Analisis dan Implementasi Long Short Term Memory Neural Network untuk Prediksi Harga Bitcoin. *E-Proceeding of Engineering Vol.5 No.2*, *5*(2), 3548–3555.

Hartono, J. (2017). *Aplikasi dan Analisis Literatur Fasilkom UI*. 4–25.

Henderi, H., & Wanda, R. L. (2017). Preprocessing Data Untuk Sistem Peramalan Tingkat Kedisiplinan Mahasiswa. *ICIT Journal*, *3*(2), 296–308. https://doi.org/10.33050/icit.v3i2.70

Kadir, A. A. W. (2021). *PERBANDINGAN KINERJA KLASIFIKASI CNN BERDASARKAN STRATEGI SPLIT DATA PADA BERAGAM DATASET CITRA*.

Kholifatullah, B. A. H., & Prihanto, A. (2023). Penerapan Metode Long Short Term Memory Untuk Klasifikasi Pada Hate Speech. *Journal of Informatics and Computer Science (JINACS)*, *04*, 292–297. https://doi.org/10.26740/jinacs.v4n03.p292-297

Mardia, E., Aisha, D., & Dimala, C. P. (2023). *Kematangan Emosi dengan Perilaku Ujaran Kebencian Pada Remaja Akhir*. *11*(2), 254–260.

Murni, M., Riadi, I., & Fadlil, A. (2023). Analisis Sentimen HateSpeech pada Pengguna Layanan Twitter dengan Metode Naïve Bayes Classifier (NBC). *JURIKOM (Jurnal Riset Komputer)*, *10*(2), 566. https://doi.org/10.30865/jurikom.v10i2.5984

Pradana, Y. A., Cholissodin, I., & ... (2023). Analisis Sentimen Pemindahan Ibu Kota Indonesia pada Media Sosial Twitter menggunakan Metode LSTM dan Word2Vec. *... Teknologi Informasi Dan ...*, *7*(5), 2389–2397.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi menggunakan MATLAB* (Nikodemus, Ed.). ANDI.

Pratama, C. H., & Findawati, Y. (2020). *Hate Speech and Emotions Classification in Indonesian Language Texts on Twitter Using Naïve Bayes Classifier [Klasifikasi Hate Speech dan Emosi Dalam Teks Berbahasa Indonesia Pada Pengguna Twitter Menggunakan Metode Naïve Bayes Classifier]*. 1–6.

Radliya, N. R. (2015). *Data mining*. *321*, 2005.

Rafael, R. Y., & Adikara, F. (2023). Pengimplmentasian Algoritma Long Short-Term Memory Untuk Mendeteksi Ujaran Kebencian Pada Aplikasi Twitter. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *8*(2), 551–560. https://doi.org/10.29100/jipi.v8i2.3490

Saputri, I. S. Y., Fadli, M., & Surya, I. (2017). Implementasi E-Commerce Menggunakan Metode UCD (User Centered Design) Berbasis Web. *Jurnal Aksara Komputer Terapan*, *6*(2), 269–278.

Wati, N. P. S., & Pramartha, C. (2022). Penerapan Long Short Term Memory dalam Mengklasifikasi Jenis Ujaran Kebencian pada Tweet Bahasa Indonesia. *Jurnal Nasional Teknologi Informasi Dan Aplikasinya (JNATIA)*, *1*(1), 755–762.

ConflictofInterestStatement: Theauthorsdeclarethattheresearchwascondu cted in the absence of any commercial or financial relationships that could be construed as a potential conflict ofinterest.