P-ISSN :2580-4316
E-ISSN :2654-8054 https://journals.upi-yai.ac.id/index.php/ikraith-informatika/issue/archive

ANALISIS SENTIMEN PADA TWITTER TERHADAP GAGALNYA PELAKSANAAN PIALA DUNIA DI INDONESIA MENGGUANAKAN METODE NAÏVE BAYES

Imam[1], Imam Santoso [2],

[1]Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, [2]Universitas Teknologi Muhammadiyah Jakarta
E-mail : imam@stikomcki.ac.id[1], imam.santoso@utmj.ac.id[2]

ABSTRAK

Analisis sentimen adalah proses mengkategorikan teks ke dalam kategori sentimen, seperti positif atau negatif. Sentimen adalah pendapat atau penilaian seseorang tentang suatu topik. Twitter adalah salah satu media sosial yang paling sering digunakan untuk berbagi pendapat tentang berbagai hal, seperti piala dunia. Tujuan penelitian ini adalah untuk mempelajari perasaan orang Indonesia tentang piala dunia, terutama terkait dengan kegagalan Indonesia untuk menjadi tuan rumah piala dunia 2022. Penelitian ini menggunakan metode naive bayes untuk mengklasifikasikan polaritas sentimen dalam tweet berbahasa Indonesia yang dihasilkan dari penerjemahan otomatis dari tweet berbahasa Inggris. Data yang digunakan dalam penelitian ini terdiri dari tweet yang mengandung kata kunci "piala dunia" dan "Indonesia" dan diposting selama periode waktu tertentu.

Untuk meningkatkan akurasi klasifikasi, penelitian ini juga menggunakan teknik pra-pemrosesan data, seleksi fitur, dan informasi nilai fitur. Studi menunjukkan bahwa meskipun Indonesia gagal menjadi tuan rumah piala dunia 2022, sebagian besar orang di Indonesia tetap memiliki perasaan positif terhadap acara tersebut. Metode naive bayes dapat mengklasifikasikan polaritas sentimen dari tweet dengan akurasi tertinggi sebesar 76.05%. Hasil penelitian ini dapat membantu mengetahui bagaimana masyarakat Indonesia melihat piala dunia. Mereka juga dapat menjadi inspirasi bagi peneliti lain untuk melakukan penelitian serupa dengan data, metode, dan teknik yang berbeda atau lebih maju.

Kata kunci : analisis sentimen, naive bayes, twitter, piala dunia, world cup, fifa.

ABSTRACT

The sentiment is a judgment or assessment of someone on a subject. Text is categorized into sentiment categories, such as Positive or negative, through sentiment analysis. One of the social media platforms that is frequently used to voice opinions on a variety of subjects, including the world cup, is Twitter. This study intends to examine Indonesians' attitudes regarding the world cup, particularly in light of Indonesia's inability to host the 2023 tournament. The naive Bayes method is used in this work to automatically categorize the polarity of sentiment in Indonesian tweets that were translated from English tweets.

The "world cup" and "Indonesia" keywords were found in tweets that were published throughout the period of this investigation. This study also uses feature selection, feature value information, and data preprocessing methods to increase classification accuracy.

The findings indicate that despite Indonesia's failure to host the 2023 World Cup, most Indonesians still favor the tournament. The polarity of sentiment from tweets can be classified with the highest accuracy of 76.05% using the naive Bayes approach.

Keyword : analisis sentimen, naive bayes, twitter, piala dunia,world cup, fifa.

P-ISSN :2580-4316
E-ISSN :2654-8054

1. PENDAHULUAN

sentimen adalah studi komputasional tentang opini, sentimen, dan emosi yang diekspresikan dalam teks. Tugas utama dari analisis sentimen adalah mengklasifikasikan polaritas dari teks yang ada dalam dokumen, kalimat, atau opini; polaritas memiliki makna jika ada teks dalam dokumen, kalimat, atau opini yang memiliki aspek positif atau negatif. Analisis sentimen dapat digunakan untuk berbagai tujuan, seperti untuk mengetahui preferensi pelanggan dan mengevaluasi kinerja produk.

Twitter adalah salah satu media sosial yang paling banyak digunakan oleh masyarakat untuk menyampaikan opini mereka; itu adalah layanan jejaring sosial mikroblogging yang memungkinkan pengguna mengirim dan membaca pesan singkat berupa teks hingga 280 karakter yang disebut tweet, yang dihasilkan oleh lebih dari 300 juta pengguna aktif setiap bulan dan menghasilkan lebih dari 500 juta tweet setiap hari. Tweet-tweet ini dapat menjadi sumber informasi yang berharga untuk mengelola informasi yang telah dikumpulkan.

Kegagalan Indonesia untuk menyelenggarakan piala dunia adalah masalah yang menarik perhatian publik. Piala dunia adalah kompetisi sepak bola internasional yang diadakan setiap empat tahun sekali oleh FIFA. Indonesia pernah menjadi tuan rumah piala dunia 2022, tetapi gagal karena korupsi, masalah keamanan, dan infrastruktur yang buruk. Setelah kegagalan ini, banyak orang di Indonesia berbicara tentang pencalonan Indonesia sebagai tuan rumah piala dunia.

Dengan menggunakan metode naive bayes pada Twitter, penelitian ini bertujuan untuk menganalisis perasaan masyarakat Indonesia tentang kegagalan menyelenggarakan piala dunia di negara mereka. Penelitian ini akan memberikan gambaran tentang bagaimana masyarakat Indonesia melihat masalah tersebut dan menawarkan masukan kepada pihak-pihak terkait untuk meningkatkan kualitas penyelenggaraan sepak bola di Indonesia. Selain itu, penelitian ini diharapkan dapat membantu kemajuan dalam bidang analisis sentimen.

Naive Bayes adalah salah satu metode klasifikasi teks yang populer dan efektif untuk analisis sentimen. Metode ini didasarkan pada teorema Bayes, yang menyatakan bahwa probabilitas suatu kelas (positif, negatif, atau netral) terhadap teks dapat dihitung dengan menggunakan probabilitas prior kelas dan probabilitas kondisional fitur (kata-kata) terhadap kelas. Metode ini dianggap naif karena menganggap bahwa fitur-fitur teks tidak bergantung satu sama lain. Namun, metode ini memiliki kelebihan karena mudah, cepat, dan akurat.

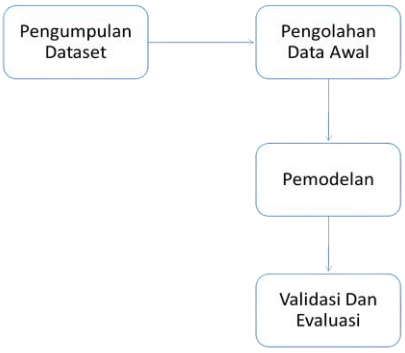
2. METODOLOGI

Studi ini adalah jenis penelitian kuantitatif yang dirancang dengan desain penelitian deskriptif. Berdasarkan data tweet yang diklasifikasikan menggunakan metode naive bayes, penelitian ini bertujuan untuk menjelaskan perasaan masyarakat Indonesia tentang kegagalan menyelenggarakan piala dunia di negara tersebut.

Tweet yang mengandung kata kunci "piala dunia" dan "Indonesia" yang diposting selama periode waktu tertentu adalah populasi penelitian ini. Penelitian ini menggunakan API Twitter untuk memilih 501 tweet secara acak dari populasi.

A.PENGUMPULAN DAN PENGOLAHAN DATA

Dalam penelitian ini, metode pengumpulan data yang digunakan adalah dokumentasi. Ini berarti mengumpulkan data dengan menggunakan sumber tertulis atau elektronik yang relevan dengan masalah penelitian. Tanggal dan waktu, teks tweet, nama pengguna, lokasi, dan informasi lainnya disimpan dalam format JSON (JavaScript Object Notation) ketika data tweet dikumpulkan dari API Twitter.



gambar 1. Metode Penelitian

Setelah melakukan pengumpulan data, dilakukan pengolahan data awal atau preprocessing. Selama tahap ini, operasi pembangunan data dan kegiatan pembersihan data yang sedang berlangsung dilakukan untuk mempersiapkan data untuk diproses. Langkah-langkah dalam persiapan data adalah sebagai berikut, termasuk

1. Stopword Removal

   Membuang kata-kata yang diabaikan pada sentiment analisis, biasanya yang berupa kata sambung dan kata keterangan.

2. Lemmatizer

   Lemmatization adalah proses pengelompokan bentuk infleksi yang berbeda dari kata yang sama. Proses ini digunakan dalam linguistik komputasi, pemrosesan bahasa alami (NLP), dan chatbot.

3. Tokenize

   Memecah sekumpulan karakter atau kalimat menjadi sebuah potongan karakter atau kata – kata sesuai dengan kebutuhan, biasa juga disebut tokenisasi.

4. Transforms Cases.

   Mengubah huruf kapital yang masih ada di dataset menjadi huruf - huruf kecil. Hal ini bertujuan agar terjadi keseragaman text pada model klasifikasi dan tidak terjadi kesalahan pada proses tokenize.

5. Indonesia Stemming

   Proses yang berfungsi mengubah kata perkata menjadi sebuah kata dasar, dengan cara menghilangkan imbuhan baik awalan maupun akhiran

6. Filter Tokens (By Lenght).

   Menghilangkan kata – kata dengan panjang karakter tertentu, biasanya kata yang memiliki hanya 2 karakter tidak memiliki arti.

Preprocessing pertama kali dilakukan secara otomatis menggunakan phyton melalui Google Collab dan google sheets. Pengolahan dataset selanjutnya dilakukan melalui tools Rapidminer.

Pada langkah pengujian metode, peneliti menggunakan RapidMiner untuk mengumpulkan data pelatihan. Data ini digunakan untuk menentukan perasaan masyarakat Indonesia tentang kegagalan menyelenggarakan piala dunia di negara mereka. Data ini kemudian dikelompokkan menjadi dua kelompok, masing-masing opini positif dan negatif.

Untuk mengevaluasi akurasi Naïve Bayes, peneliti menggunakan 10 Fold Cross Validation.

sebagai garis horizontal dan true positif sebagai garis vertikal. Pedoman umum untuk mengklasifikasikan keakuratan pengujian menggunakan AUC [5].

0.90 - 1.00 = Excellent Classification;

0.80 - 0.90 = Good Classification;

0.70 - 0.80 = Fair Classification;

0.60 - 0.70 = Poor Classification;

0.50 - 0.60 = Failure

P-ISSN :2580-4316
E-ISSN :2654-8054

3. LANDASAN TEORI

A. Analisis Sentimen

Analisis sentimen, juga dikenal sebagai Opinion Mining, adalah bidang studi yang bertujuan untuk menganalisis opini, sentimen, penilaian, evaluasi, sikap, dan emosi publik terhadap suatu entitas dari produk, layanan, masalah, organisasi, peristiwa, atau fitur tertentu. Analisis sentimen dapat dilakukan dengan berbagai metode, seperti NB (Naive Bayes), Decision Tree, KNN (K-Nearest Neighbor), Neural Networks, dan SVM (Support Vector Machines).

B. Rapid Miner

RapidMiner adalah perangakat lunak open-source. RapidMiner menawarkan analisis data mining, text mining, dan prediksi. Berbagai teknik deskriptif dan prediksi membantu pengguna membuat keputusan yang lebih baik. RapidMiner adalah software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri, dengan lebih dari 500 operator untuk input, output, data preprocessing, dan visualisasi. RapidMiner berjalan di semua sistem operasi karena dibuat dengan bahasa Java.

RapidMiner awalnya dikenal sebagai YALE (Yet Another Learning Environment), dan dimulai pada tahun 2001 oleh AlfKlinkenberg, Ingo Mierswa, dan Simon Fischer bekerja di Unit Kecerdasan Buatan Universitas Dortmund. Saat ini, ribuan aplikasi yang dikembangkan menggunakan RapidMiner di lebih dari empat puluh negara, dan itu tersedia di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Karena popularitasnya yang luar biasa di seluruh dunia, RapidMiner jelas merupakan software data mining open source.

Menurut polling yang dilakukan oleh KDnuggets, sebuah portal data mining, RapidMiner adalah software data mining terbaik. Pada tahun 2010–2011, GUI (Graphic User Interface) digunakan untuk merancang pipeline analitis dan menghasilkan file XML (Extensible Markup Language) yang menjelaskan proses analitis yang diinginkan pengguna untuk diterapkan ke data. File ini kemudian dibaca oleh RapidMiner untuk menjalankan analis secara otomatis.

Fitur Rapidminer diantara lain :

1. Banyaknya algoritma data mining, seperti decision tree dan self organization map.
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.
3. Banyaknya variasi plugin, seperti text plugin untuk melakukan analisis teks.
4. Menyediakan prosedur data mining dan machine learning termasuk: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi
5. Proses data mining tersusun atas operator - operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI.

C. Naive bayes adalah salah satu metode klasifikasi teks yang populer dan efektif untuk analisis sentimen. Metode ini berdasarkan pada teorema Bayes yang menyatakan bahwa probabilitas suatu kelas (positif, negatif, atau netral) terhadap suatu teks dapat dihitung dengan menggunakan probabilitas prior kelas dan probabilitas kondisional fitur (kata-kata) terhadap kelas. Metode ini disebut naive karena mengasumsikan bahwa fitur-fitur dalam teks bersifat independen satu sama lain. Meskipun demikian, metode ini memiliki kelebihan dalam hal simpel, cepat, dan akurat.

4. HASIL DAN PEMBAHASAN

A. HASIL ANALISIS DATA

Berdasarkan hasil analisis data yang dilakukan pada bab sebelumnya, beberapa hal dapat disimpulkan sebagai berikut:

Dari 501 tweet sampel, 434 berlabel positif dan 67 berlabel negatif. Hal ini menunjukkan bahwa meskipun Indonesia gagal menjadi tuan rumah piala dunia U-20 2023, sebagian besar orang di Indonesia masih menyukai piala dunia. Metode naive Bayes memiliki akurasi tertinggi sebesar 76.05% dalam mengklasifikasikan polaritas sentimen dari tweet.

Hal ini menunjukkan bahwa bagian terakhir dokumen mengandung lebih banyak informasi sentimen yang relevan daripada bagian awal atau tengah. Oleh karena itu, penggunaan fitur dari 25% bagian terakhir dokumen menghasilkan hasil yang lebih baik

daripada penggunaan fitur dari keseluruhan dokumen.

B. Pembahasan

Hasil penelitian ini sesuai dengan beberapa penelitian sebelumnya yang menggunakan metode naive bayes untuk analisis sentimen pada dokumen berbahasa Inggris maupun Indonesia, seperti:

Penelitian Franky (2008) yang menggunakan metode naive bayes untuk analisis sentimen pada dokumen review film berbahasa Inggris dan dokumen review film berbahasa Indonesia hasil penerjemahan otomatis. Penelitian ini mendapatkan akurasi tertinggi sebesar 80.09% pada dokumen berbahasa Inggris dan 78.82% pada dokumen berbahasa Indonesia.

Penelitian Sari et al. (2019) yang menggunakan metode naive bayes untuk analisis sentimen pada dokumen review produk online berbahasa Indonesia. Penelitian ini mendapatkan akurasi tertinggi sebesar 86% dengan menggunakan seleksi fitur chi square feature selection dan informasi nilai fitur presence.

Penelitian Pratama et al. (2020) yang menggunakan metode naive bayes untuk analisis sentimen pada dokumen review film berbahasa Indonesia. Penelitian ini mendapatkan akurasi tertinggi sebesar 84% dengan menggunakan seleksi fitur information gain dan informasi nilai fitur presence.

Hasil penelitian ini juga konsisten dengan beberapa teori yang berkaitan dengan analisis sentimen, seperti:

Teori Liu (2012) yang menyatakan bahwa analisis sentimen adalah studi komputasional tentang opini, sentimen, dan emosi yang diekspresikan dalam teks, dan tugas dasarnya adalah mengklasifikasikan polaritas dari teks yang ada dalam dokumen, kalimat, atau opini. Penelitian ini menggunakan teks berupa tweet sebagai sumber data dan mengklasifikasikan polaritasnya menjadi positif atau negatif.

Teori Pang et al. (2002) yang menyatakan bahwa metode naive bayes adalah salah satu metode klasifikasi teks yang populer dan efektif untuk analisis sentimen. Penelitian ini menggunakan metode naive bayes untuk menghitung probabilitas posterior kelas terhadap teks dengan menggunakan probabilitas prior kelas dan probabilitas kondisional fitur terhadap kelas.

Teori Han & Kamber (2006) yang menyatakan bahwa seleksi fitur adalah proses memilih sebagian fitur dari keseluruhan fitur yang tersedia untuk digunakan dalam klasifikasi. Penelitian ini menggunakan seleksi fitur chi square feature selection untuk memilih fitur-fitur yang paling relevan untuk klasifikasi.

5. KESIMPULAN

Berdasarkan hasil analisis data yang dilakukan pada bab sebelumnya, beberapa hal dapat disimpulkan sebagai berikut:

Dari 501 tweet sampel, 434 berlabel positif dan 67 berlabel negatif. Hal ini menunjukkan bahwa meskipun Indonesia gagal menjadi tuan rumah piala dunia U-20 2023, sebagian besar orang di Indonesia masih menyukai piala dunia. Metode naive Bayes memiliki akurasi tertinggi sebesar 76.05% dalam mengklasifikasikan polaritas sentimen dari tweet.

Hal ini menunjukkan bahwa bagian terakhir dokumen mengandung lebih banyak informasi sentimen yang relevan daripada bagian awal atau tengah. Oleh karena itu, penggunaan fitur dari 25% bagian terakhir dokumen menghasilkan hasil yang lebih baik daripada penggunaan fitur dari keseluruhan dokumen.

P-ISSN :2580-4316
E-ISSN :2654-8054

(Rekayasa Sistem dan Teknologi Informasi), 3(3), 379-385.

DAFTAR PUSTAKA

1. Franky. (2008). Analisis Sentimen Menggunakan Metode Naive Bayes, Maximum Entropy, dan Support Vector Machine pada Dokumen Berbahasa Inggris dan Dokumen Berbahasa Indonesia Hasil Penerjemahan Otomatis. Skripsi. Fakultas Ilmu Komputer, Universitas Indonesia. Diakses dari https://lib.ui.ac.id/file?file=digital/123194-SK-705-Analisis%20sentimen-HA.pdf
   i.
2. Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann.
   i.
3. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.
   i.
4. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.
   i.
5. Pratama, A., Wibowo, A., & Suryana, N. (2020). Analisis Sentimen Review Film Menggunakan Metode Naive Bayes Classifier dengan Seleksi Fitur Information Gain. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 4(1), 1-7.
   i.
6. Sari, D., Wibowo, A., & Suryana, N. (2019). Analisis Sentimen Review Produk Online Menggunakan Metode Naive Bayes Classifier dengan Seleksi Fitur Chi Square. Jurnal RESTI