Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

Klasifikasi Sentimen Komentar Youtube Tentang Pembatalan Indonesia Sebagai Tuan Rumah Piala Dunia U-20 Menggunakan Algoritma Naïve Bayes Classifer

Ilham Habibi Hasibuan, Elvia Budianita, Surya Agustian, Pizaini

Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]111950115079@students.uin-suska.ac.id, [2,*]elvia.budianita@uin-suska.ac.id, [3]surya.agustian@uin-suska.ac.id, [4]pizaini@uin-suska.ac.id
Email Penulis Korespondensi: elvia.budianita@uin-suska.ac.id
Submitted: **07/12/2023**; Accepted: **20/12/2023**; Published: **22/12/2023**

Abstrak–Text mining adalah metode yang digunakan untuk melakukan tugas-tugas seperti klasifikasi dokumen, pengelompokan, ekstraksi informasi, analisis sentimen, dan pengambilan informasi. Federation Internationale Football Association (FIFA), badan pengatur sepak bola internasional, telah menetapkan Indonesia sebagai negara tuan rumah Piala Dunia U-20 mulai tahun 2019. Indonesia diperkirakan akan menjadi pilihan venue Piala Dunia U-20 pada tahun 2021. Namun, akibat wabah Covid-19, Piala Dunia dijadwal ulang dan kini dijadwalkan berlangsung pada tahun 2023. Indonesia secara resmi melepaskan posisinya sebagai tuan rumah pada 31 Maret 2023. Salah satu alasannya adalah banyaknya fraksi yang menentang kehadiran timnas Israel di Indonesia. Alhasil, beragam reaksi masyarakat menyikapi keputusan Indonesia yang membatalkan penyelenggaraan Piala Dunia U-20, khususnya pada video channel YouTube Narasi tv bertajuk "Piala Dunia U-20 Gagal Digelar di Indonesia, Mari Kita Lihat dari Dua Perspektif | Musyawarah". Sejak video tersebut diunggah hingga 16 Agustus 2023, total komentar yang dihasilkan sebanyak 4.629 komentar. Penelitian ini menggunakan pendekatan Naïve Bayes classifier. Naïve Bayes Classifier (NBC) adalah pengklasifikasi probabilistik langsung yang memanfaatkan Teorema Bayes dalam kondisi independensi yang kuat. Pengujian yang dilakukan menunjukkan bahwa performa model ketika menggunakan teknik stopword removal dan stemming lebih unggul dalam mengklasifikasi kelas dalam dataset. F1-Score sebesar 59,70% dan nilai Accuracy sebesar 63,43%. Selanjutnya, setelah mengidentifikasi model yang paling efisien untuk menerapkan klasifikasi naïve Bayes, evaluasi dilakukan pada data validasi menghasilkan hasil F1-Score sebesar 58,72% dan tingkat akurasi sebesar 61,65%. Analisis klasifikasi menunjukkan bahwa masyarakat Indonesia mempunyai pandangan negatif atau mengalami kekecewaan atas pembatalan tersebut.

Kata Kunci: Analisis Sentimen; FIFA; Naïve Bayes Classifer; Piala Dunia U-20; Youtube

Abstrak–Text mining is a method used to perform tasks such as document classification, clustering, information extraction, sentiment analysis, and information retrieval. The Federation Internationale Football Association (FIFA), the international football governing body, has designated Indonesia as the host country for the U-20 World Cup starting in 2019. Indonesia is expected to be the choice venue for the U-20 World Cup in 2021. However, due to the Covid outbreak -19, the World Cup was rescheduled and is now scheduled to take place in 2023. Indonesia officially relinquished its position as host on March 31 2023. One of the reasons is the many factions that oppose the presence of the Israeli national team in Indonesia. As a result, various public reactions responded to Indonesia's decision to cancel holding the U-20 World Cup, especially on the Narasi tv YouTube channel video entitled "The U-20 World Cup Failed to Be Held in Indonesia, Let's Look at it from Two Perspectives | Discussion". Since the video was uploaded until August 16 2023, the total comments generated were 4,629 comments. This research uses a Naïve Bayes classifier approach. Naïve Bayes Classifier (NBC) is a direct probabilistic classifier that exploits Bayes' Theorem under strong independence conditions. The tests carried out show that the model performance when using stopword removal and stemming techniques is superior in classifying classes in the dataset. The F1-Score is 59.70% and the Accuracy value is 63.43%. Furthermore, after identifying the most efficient model for applying naïve Bayes classification, evaluation was carried out on validation data resulting in an F1-Score of 58.72% and an accuracy rate of 61.65%. Classification analysis shows that Indonesian people have a negative view or are disappointed with the cancellation.

Keywords: Sentiment Analysis; FIFA; Naïve Bayes Classifier; U-20 World Cup; YouTube

1. PENDAHULUAN

Perkembangan teknologi informasi mempunyai pengaruh yang sangat besar terhadap setiap aspek kehidupan sehari-hari. Perkembangan tersebut tidak terlepas dengan penyebaran informasi kepada publik dengan media internet, dimana dengan menggunakan media internet di perkirakan lebih efisien dan efektif dalam penyampaian informasi. Akan tetapi, penyebaran informasi yang begitu cepat dan dapat di akses oleh semua orang menimbulkan berbagai reaksi terhadap informasi yang diterima. Information Retrieval (IR) adalah pendekatan sistematis untuk menstrukturkan, merepresentasikan, menyimpan, dan mencari informasi dalam berbagai format seperti teks dan multimedia [1]. Text Mining terus menghadapi tantangan yang signifikan, dengan Pengambilan Informasi menjadi salah satu tugas tersulit di bidang ini [2].

Text Mining adalah metode yang digunakan untuk melakukan tugas-tugas seperti klasifikasi dokumen, pengelompokan, ekstraksi informasi, analisis sentimen, dan pengambilan informasi. Ini adalah bentuk khusus penambangan data yang bertujuan untuk mengidentifikasi pola signifikan dalam kumpulan data tekstual yang luas [3]. Text Mining, atau penambangan data teks, adalah analisis informasi berharga dengan memeriksa pola dan tren statistic [4]. Analisis sentimen adalah teknik yang digunakan untuk mengumpulkan data opini dan secara otomatis menganalisis materi tekstual untuk menentukan sentimen yang diungkapkan dalam suatu opini.[5]

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 249
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

Federation Internationale Football Association (FIFA), badan pengatur sepak bola internasional, telah menetapkan Indonesia sebagai negara tuan rumah Piala Dunia U-20 mulai tahun 2019. Indonesia diperkirakan akan menjadi pilihan venue Piala Dunia U-20 pada tahun ini. 2021. Piala Dunia U-20 yang dijadwalkan pada tahun 2021 diundur ke tahun 2023 karena pandemi Covid-19 yang masih berlangsung. Posisi Indonesia sebagai tuan rumah Piala Dunia U-20 pada 31 Maret 2023 terancam batal akibat keberatan Gubernur Provinsi Bali I Wayan Coster terhadap keikutsertaan Timnas Israel.[6].

Potensi batalnya peran Indonesia sebagai tuan rumah Piala Dunia U-20 tentu menimbulkan rasa kecewa di kalangan individu, karena turnamen yang sudah mendunia ini menjanjikan banyak manfaat bagi berbagai lapisan masyarakat. Keputusan FIFA menunda pengundian Piala Dunia U-20 yang semula dijadwalkan pada 31 Maret di Denpasar, Bali, memunculkan isu pembatalan. Keputusan ini lekat dengan perselisihan tersingkirnya Timnas Israel yang sebelumnya meraih peringkat kedua Grup B Piala Eropa U-19 2022. Banyak orang memanfaatkan saluran media sosial mereka yang berbeda untuk berbagi emosi yang mereka alami. Platform media sosial seperti Twitter, Instagram, YouTube, dan Facebook berfungsi sebagai saluran untuk berbagi informasi dan memfasilitasi diskusi interaktif. Salah satu media yang mengangkat topik tersebut menjadi bahan tang adalah channel youtube Narasi.tv dengan judul "Piala Dunia U-20 gagal digelar di Indonesia, Mari Lihat dari Dua Perspektif | Musyawarah " mendapatkan komentar sebayak 4629 komentar sejak video tersebut di upload sampai dengan tanggal 16 Agustus 2023.

Penelitian yang dilakukan oleh Fransiska Vina Sari dan Arif Wibowo pada tahun 2019 berjudul "Analisis Sentimen Pelanggan Toko Online Jd.Id". Temuan penelitian menunjukkan bahwa analisis sentimen adalah metode penggalian data tekstual untuk memperoleh informasi terkait sentimen, seperti nilai positif, netral, atau negatif, menggunakan Metode Naïve Bayes Classifier berdasarkan Emotional Icon Conversion. Pengguna internet di media sosial melakukan analisis sentimen untuk memberikan evaluasi atau sudut pandang subjektif mereka. Temuan penelitian menunjukkan bahwa pendekatan Naïve Bayes, bila digunakan tanpa fitur tambahan, mencapai akurasi klasifikasi sentimen sebesar 96,44%. Namun, ketika fitur pembobotan tf-idf digabungkan dengan konversi ikon emosional, nilai akurasinya meningkat menjadi 98% [5].

Kajian bertajuk "Analisis Sentimen Respon Masyarakat terhadap Capres 2024 Ridwan Kamil" ini dilakukan oleh Neni Sari Putri Juana, Elin Haerani, Fadhilah Syafria, dan Elvia Budianita pada tahun 2023 memberikan hasil penelitian menggunakan Metode Naive Bayes Classifier untuk menerapkan dan menguji metode Naïve Bayes pada tiga skenario sebaran data (train:test), yaitu 90%:10, 80%:20%, dan 70%:30%. Skor akurasi tertinggi sebesar 86,38% dicapai pada ketiga skenario ini, dengan menggunakan partisi data 80%:20%. Dataset terdiri dari 1262 komentar positif dan 242 komentar negatif [7].

Kajian bertajuk "Analisis Sentimen Aplikasi Ruangguru" ini dilakukan pada tahun 2020 oleh Evita Fitri, Yuri Yuliani, Susy Rosyida, dan Windu Gata. Penerapan Algoritma Naive Bayes, Random Forest, dan Support Vector Machine menunjukkan bahwa eksperimen yang dilakukan menghasilkan tingkat akurasi yang sangat tinggi dibandingkan dengan metode lain. Algoritma klasifikasi Naive Bayes mengalami peningkatan sebesar 1,85% dibandingkan hasil sebelumnya, dengan akurasi sebesar 96,01%. Nilai presisi untuk prediksi benar dan prediksi salah masing-masing adalah 100,00% dan 92,60%, namun nilai recall untuk setiap prediksi tidak ditentukan. senilai 92,01% dan 100,00% [8].

Pada tahun 2021, Yuyun, Nurul Hidayah, dan Supriadi Sahibu menghasilkan penelitian bertajuk "Algoritma Multinomial Naïve Bayes untuk Klasifikasi Sentimen Pemerintah Terkait Penanganan Covid-19". Pengujian algoritma Multinomial Naïve Bayes di Twitter Data mengungkapkan bahwa model tersebut mencapai rata-rata akurasi, presisi, dan recall masing-masing sebesar 74% dalam memprediksi sentimen terhadap penanganan Covid di Indonesia. Karena penggunaan query yang berurutan dalam pencarian dokumen teks, hasil pengujian untuk ketiga parameter yang disebutkan di atas menunjukkan nilai yang identic [9].

Kajian yang dilakukan pada tahun 2023 oleh Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, dan Fitri Nurapriani bertajuk "Analisis Sentimen Pemindahan Ibukota Kepulauan". Pemanfaatan algoritma Naïve Bayes dan KNN menghasilkan temuan kajian yang memungkinkan dilakukannya penentuan berbagai tahapan dalam menganalisis data sentimen Twitter terkait kemajuan Ibu Kota Negara Nusantara. Tahapan tersebut meliputi: Data tweet dianalisis menggunakan metode Naive Bayes (NB) untuk analisis sentimen. Tingkat akurasi analisis sebesar 82,27%, dengan nilai presisi sebesar 86,36% dan nilai recall sebesar 76,93%. Selain itu, metode K-Nearest Neighbors (KNN) memperoleh tingkat akurasi sebesar 88,12%, presisi sebesar 93,98%, dan nilai recall sebesar 81,53% [10]

Kajian bertajuk "Analisis Sentimen Opini Publik Terhadap Film di Platform Twitter" ini dilakukan pada tahun 2022 oleh Yuni Nurtikasari, Syariful Alam, dan Teguh Iman Hermanto. Memanfaatkan Algoritma Naive Bayes, penelitian analisis sentimen pada film Ngeri-ngeri Sedap dilakukan pada platform Twitter dengan total 404 titik data. Teks telah melalui proses awal yang meliputi penyaringan, tokenisasi, transformasi, dan klasifikasi menggunakan metode Naïve Bayes. Selain itu, evaluasi data telah dilakukan menggunakan matriks konfusi dengan bantuan alat Orange. Hasil analisis menunjukkan bahwa sentimen masyarakat terhadap film Ngeri-ngeri Sedap didominasi netral dengan tingkat akurasi sebesar 75%. Penilaian ini didasarkan pada nilai presisi sebesar 80% dan tingkat keberhasilan (recall) sebesar 79%. Data tersebut menunjukkan bahwa mood populer terhadap film Ngeri-ngeri Sedap di jaringan Twitter termasuk dalam kategori netral.[11]

Melihat permasalahan di atas, maka dilakukan penelitian untuk mengkaji emosi masyarakat Indonesia atas

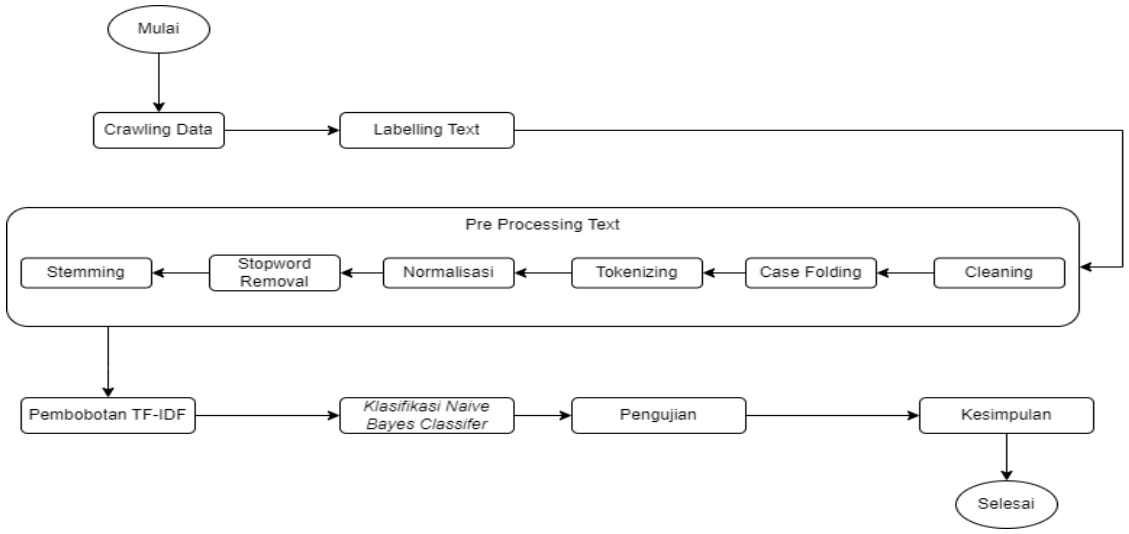
Copyright © 2023 **Ilham Habibi Hasibuan**, Page 250
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

keputusan mundurnya tuan rumah Piala Dunia U-20. Algoritma Naïve Bayes Classifier diterapkan untuk menganalisis skenario tersingkirnya Indonesia sebagai tuan rumah Piala Dunia U-20 pada penelitian ini. Tujuan dari penelitian ini adalah untuk mengetahui sentimen atau reaksi masyarakat Indonesia terhadap pembatalan tersebut, sekaligus mengevaluasi efektivitas algoritma Naïve Bayes Classifier dalam mengkategorikan sentimen secara akurat.

2. METODOLOGI PENELITIAN

Penelitian ini bertujuan untuk menggunakan pendekatan klasifikasi Naïve Bayes Classifier untuk melakukan klasifikasi sentimen seputar pembatalan Indonesia sebagai tuan rumah Piala Dunia U 20. Penelitian akan berkembang melalui langkah-langkah yang diuraikan pada Gambar 1 di bawah.



Gambar 1. Desain Penelitian

Penelitian ini meliputi beberapa tahap yaitu crawling data, pelabelan teks, preprocessing, pembobotan TF-IDF, klasifikasi menggunakan Naïve Bayes Classifier, pengujian, dan penarikan kesimpulan. Paragraf berikutnya memberikan penjelasan komprehensif tentang langkah-langkah tersebut.

2.1 Crawling Data

Crawling data adalah proses yang pertama kali dilakukan dalam text mining. Crawling data (pengumpulan data) dapat bersumber dari beberapa platform baik media sosial, media berita, rating dan lain sebagainya. Proses ini dilakukan menggunakan library python, data yang didapat berbentuk file json kemudian diubah menjadi file excel.[12]

2.2 Labelling Text

Labelling text (pelabelan) adalah Proses penetapan kelas pada data (komentar) yang dihasilkan melalui proses perayapan. Komentar tersebut akan dibagi menjadi tiga kelas, yaitu kelas positif, negatif, dan netral. Proses pelabelan pada penelitian ini menggunakan metode crowdsourcing yang melibatkan lebih dari satu anotator. Hal ini dilakukan untuk menghilangkan unsur subjektif dalam permasalahan yang diteliti.[13]

2.3 Pre-processing Text

Pre-processing merupakan tahap awal yang dilakukan pada saat implementasi text mining [14]. Tujuan dari pra-pemrosesan adalah untuk memperoleh data dalam format yang dapat diinterpretasikan oleh algoritma yang akan digunakan. Pre-processing teks melibatkan beberapa fase yaitu cleaning, case folding, tokenizing, normalisasi, stopword removal dan stemming[15].

2.4 Pembobotan TF-IDF

Pendekatan Term Frekuensi Invers Dokumen Frekuensi (TF-IDF) adalah teknik yang digunakan untuk menilai relevansi kata (istilah) pada suatu dokumen dengan memberikan bobot pada setiap kata. Pendekatan TF-IDF merupakan kombinasi dari dua konsep: frekuensi kata dalam sebuah dokumen dan frekuensi dokumen kebalikan dari kata tersebut [16]. TF-IDF dihitung berdasarkan persamaan berikut.

$$W_{dt} = tf_{d.t} \text{ X } idf_{d.t} \quad (1)$$

$$IDF_t = \log\left(\frac{D}{d_{f.t}}\right) \quad (2)$$

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 251
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

Adapun keterangan dari persamaan diatas adalah sebagai berikut, Wdt atau Besarnya kemunculan suatu kata ke-t pada suatu dokumen d, TFdt atau Banyaknya sebuah term pada suatu dokumen, IDFt atau Frekuensi inverse sebuah dokumen, N adalah Jumlah keseluruhan dokumen, Df atau Jumlah dokumen yang mempunyai suatu term yang telah ditentukan.

2.5 Klasifikasi Naïve Bayes Classifer

Naïve Bayes Classifier (NBC) adalah pengklasifikasi probabilistik langsung yang menggunakan Teorema Bayes dengan premis independensi yang kuat [17]. Teknik Naïve Bayes dapat dihitung menggunakan persamaan berikut.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

Persamaan di atas dapat dijelaskan sebagai berikut: Variabel yang dimaksud adalah X yang mewakili data yang kelasnya tidak diketahui, dan H yang mewakili hipotesis data atau probabilitas hipotesis. P(X|H) menunjukkan probabilitas hipotesis X dengan kondisi H, sedangkan P(H) mewakili probabilitas sebelumnya dari hipotesis H. Terakhir, P(X) mengacu pada probabilitas hipotesis X.

2.6 Confusion Matrix

Matriks kebingungan digunakan dalam pengujian algoritma untuk menghitung nilai presisi, perolehan, dan akurasi [19]. Pendekatan ini menggunakan tabel matriks, seperti yang diilustrasikan pada tabel di bawah ini.

Tabel 1. Confusion Matriks

	Predicted Positif	**Predicted Negatif**	**Predicted Netral**
Actual Positif	TP	FN	FNet
Actual Negatif	FP	TN	FP
Actual Netral	FP	FP	TNet

Untuk menghitung nilai akurasi, presisi recall dan F1-score dapat dilihat pada persamaan berikut ini.

$$\text{Akurasi} = \frac{TP+TN+Tnet}{TP+TN+TNet+FP+FN+FNet} *100 \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} *100 \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1-Score} = 2 * \frac{precision*recall}{precision*+recall} \quad (7)$$

3. HASIL DAN PEMBAHASAN

3.1 Crawling Data

Metode pengumpulan data mengacu pada serangkaian teknik yang digunakan oleh peneliti untuk mengumpulkan data [18]. Pengumpulan data (crawling) pada penelitian kali ini menggunakan lybrary python sebagai crawler. Data yang dihasilkan berupa file dalam bentuk json kemudian dilakukan convert ke dalam bentuk excel. Adapun hasil dari crawling tersebut mengumpulkan sebanyak 4629 komentar terhitung dari video di upload sampai dengan tanggal 16 Agustus 2023. Berikut ini hasil crawling data disajkan dalam tabel 2 berikut ini.

Tabel 2. Hasil Crawling Dataset

No	**Komentar**	**Sentimen**
1	Harusnya pemerintahan perlu kandidat muda biar berpikirnya gak kolot,biar pemikiran melenial terbuka,kalau gak negara ini gak akan maju maju	Negatif
2	Kami sangat setuju pembatalan tuan rumah u20.. karena komitmen bangsa kita tidak tergoyahkan apapun yang terjadi.	Positif
3	Sebenarnya sederhana palestina (yg dibela mati2an) sama sekali TIDAK mempermasalahkan israel ikut tanding,lalu kenapa indonesia malah mempermasalahkannya?? Sampai harus mengorbankan impian garuda muda?	Netral

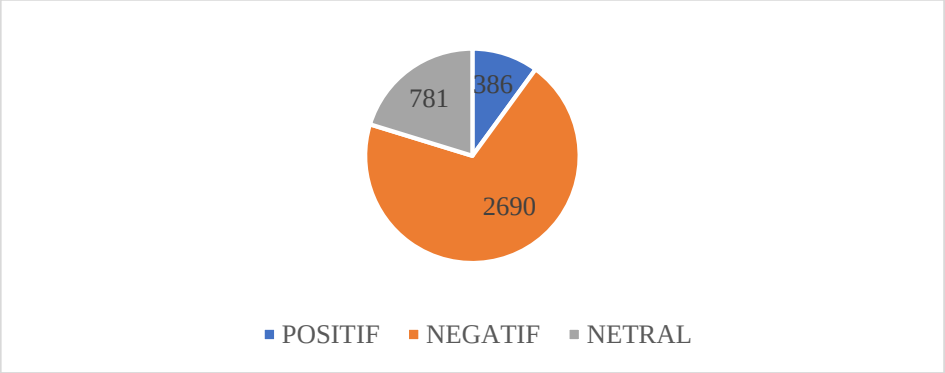
3.1 Labeling Text

Labeling[19] atau pemberian label pada setiap komentar ini dilakukan bersama beberapa mahasiswa dengan menganalisa komentar tersebut apakah masuk ke dalam kelas positif, negatif, atau netral. Dalam menentukan kelas sentiimen pada dataset dilakukan beberapa pendekatan yaitu untuk komentar yang berisikan kalimat pujian, semangat dan rasa bangga di masukkan kedalam kategori kelas positif, sementara untuk kamentar seperti ujaran

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 252
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

kebencian, rasa kecewa dan rasa marah dimasukkan kedalam kategori negatif. Sedangkan untuk menentukan kelas adalah komentar yang tidak termasuk kedalam kategori positif dan negatif. Dari data yang dianalisa menghasilkan 4629 data kemudian dilakukan proses filtering dengan menghapus data duplikat dan data yang tidak sesuai dengan kasus sehingga data yang dihasilkan setelah proses filtering sebanyak 3857 data dengan jumlah data dengan senntimen negatif sebanyak 2690 data, sentimen positif 386 data dan sentimen netral 781 data. berikut ini disajikan diagram distribusi kelas pada dataset pada gambar 2.



Gambar 2. Distribusi Sentimen dalam Dataset

3.3 Pre-processing Text

Pre Processing merupakan tahap awal yang dilakukan pada saat implementasi text mining. Tujuan dari pra-pemrosesan di Text Mining adalah untuk mengekstrak informasi berharga dari data tidak terstruktur dan menghilangkan istilah-istilah yang tidak relevan dari dokumen[14]. Pre Processing teks adalah fase awal di mana data teks termasuk noise diminimalkan untuk memfasilitasi pengurangan noise lebih lanjut selama pemrosesan. [20]. Pada tahapan pre processing dilakukan beberapa proses yaitu cleaning, case folding, tokenizing, normaliassi, stopword removal, dan stemming.

3.3.1 Cleaning

Cleaning adalah penghapusan tanda baca, angka, simbol, tautan URL, dan nama pengguna dari teks. Penghapusan stopword berarti menghilangkan kata-kata yang dianggap tidak penting di dalam teks. [21]. Pada tabel 3 disajikan hasil cleaning. Keseluruhan proses cleaning dilakukan menggunakan bahasa php dan dijalankan melalui google colab.

3.3.2 Case Folding

Tahapan case folding[14] melibatkan transformasi teks ke bentuk seragam dalam hal kapitalisasi huruf. Tujuannya atau mempermudah perbandingan dan pencarian teks tanpa memperhatikan apakah hurufnya besar atau kecil. Umumnya, case folding digunakan dalam pengolahan bahasa algoritma serta pencarian teks.

3.3.3 Tokenizing

Tokenizing adalah prosedur membagi teks atau data yang panjang menjadi komponen-komponen yang lebih kecil yang dikenal sebagai token. Token adalah unit teks yang tidak dapat dipisahkan dan memiliki makna atau representasi tertentu. Token mencakup banyak elemen seperti kata, frasa, simbol, atau komponen lainnya, yang ditentukan oleh konteks spesifik pemrosesan data [22].

3.3.4 Normalisasi

Normalisasi adalah langkah penting dalam mengidentifikasi kesalahan dalam istilah, seperti duplikat atau singkatan, yang memiliki arti yang sama. Dengan mengoreksi kata-kata yang salah ini, hal ini membantu menciptakan data yang lebih terorganisir dan efisien. [23]. Beberapa kata yang termasuk dalam daftar normalisasi adalah kata “sekarang” yang disingkan menjadi “skrg” dan kata “sebelum” yang disingkat menjadi “sblm”.

3.3.5 Stopword Removal

Stopword removal atau tahapan menghilangkan kata penghubung dalam sebuah kalimat[8] misalnya “ini”, “apa” dan lain sebagainya. Keuntungan dari penghapusan stopword meliputi pengurangan ukuran indeks, peningkatan akurasi pencarian, dan efisiensi pemrosesan. Namun, perlu diingat bahwa dalam beberapa situasi, penghapusan stopword mungkin tidak selalu diinginkan dan harus dipertimbangkan dengan cermat berdasarkan konteks dan jenis data yang dihadapii

3.3.4 Stemming

Stemming atau proses dalam pemrosesan bahasa alami yang bertujuan untuk menghapus infleksi atau akhiran kata

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 253
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

sehingga hanya menyisakan akar kata atau "stemma". Ini membantu dalam pra-pemrosesan teks, kata, dan dokumen untuk normalisasi teks. Stemming mengurangi kata-kata ke bentuk dasar mereka, yang mungkin atau mungkin bukan merupakan kata yang sah dalam bahasa. Misalnya, kata-kata “memakai” dan “dipakai” memiliki akar yang sama, yaitu “pakai”. Tujuannya adalah untuk memudahkan model dalam melakukan klasifikasi kelas pada data yang sudah dikembalikan menjadi akar kata [24]

Setelah melakukan semua proses text preprocessing selesai maka data tersebut sudah siap untuk dilakukan proses selanjutnya sampai dengan penerapan algoritma untuk menemukan hasil klasifikasi sentimen. Berikut ini disajikan pada tabel 3 hasil dari proses text preprocessing tersebut.

Tabel 3. Hasil Text Preprocessing

No	Proses	Sebelum Proses	Setelah Proses
1	Cleaning	Hahahahah... Lucunya kalian yang menolak Israel 😂😂😂 Benang merah apanya.. . Lucu-lucu itu perkataanmu itu loh	Hahahahah Lucunya kalian yang menolak Israel Benang merah apanya Lucu lucu itu perkataanmu itu loh
2	Casefolding	Hahahahah Lucunya kalian yang menolak Israel Benang merah apanya Lucu lucu itu perkataanmu itu loh	hahahahah lucunya kalian yang menolak israel benang merah apanya lucu lucu itu perkataanmu itu loh
3	Tokenizing	hahahahah lucunya kalian yang menolak israel benang merah apanya lucu lucu itu perkataanmu itu loh	'hahahahah', 'lucunya', 'kalian', 'yang', 'menolak', 'israel', 'benang', 'merah', 'apanya', 'lucu', 'lucu', 'itu', 'perkataanmu', 'itu', 'lo'
4	Normalisasi	'hahahahah', 'lucunya', 'kalian', 'yang', 'menolak', 'israel', 'benang', 'merah', 'apanya', 'lucu', 'lucu', 'itu', 'perkataanmu', 'itu', 'lo'	'hahahahah', 'lucunya', 'kalian', 'yang', 'menolak', 'israel', 'benang', 'merah', 'apanya', 'lucu', 'lucu', 'itu', 'perkataanmu', 'itu', 'lo'
5	Stopword Removal	'hahahahah', 'lucunya', 'kalian', 'yang', 'menolak', 'israel', 'benang', 'merah', 'apanya', 'lucu', 'lucu', 'itu', 'perkataanmu', 'itu', 'lo'	'hahahahah', 'lucunya',, 'menolak', 'israel', 'benang', 'merah', 'apanya', 'lucu', 'lucu', , 'perkataanmu'
6	Stemming	'hahahahah', 'lucunya', 'menolak', 'israel', 'benang', 'merah', 'apanya','lucu', 'perkataanmu'	'lucu',, 'tolak', 'israel', 'benang', 'merah', 'apa', 'lucu', kata'

3.4 Pembobotan TF-IDF

Pembobotan TF-IDF diterapkan untuk memberikan bobot atau nilai pada setiap kata. TF-IDF memungkinkan evaluasi frekuensi kata dalam dokumen (Frekuensi Jangka) dan pentingnya istilah dalam kumpulan dokumen lengkap (Frekuensi Dokumen Invers). Perkalian Term Frekuensi (TF) dan Inverse Document Frekuensi (IDF) menghasilkan skor TF-IDF, yang secara akurat mewakili signifikansi suatu istilah dalam konteks dokumen tertentu dan keseluruhan koleksi dokumen. Hasil pembobotan TF IDF dilaporkan pada tabel 8.

Tabel 4. Hasil TF- IDF

No.	perintah	negara	main	maju	bangga	tuan	politik
1	0,195495	0,128271	0	0,395482	0	0	0
2	0,503758	0	0	0	0	0	0
3	0	0	0,236874	0	0,39344	0,242498	0
…	….	…	…	….	….	….	….
3857	0	0	0	0	0	0	0,33108

3.5 Eksperimen Setup

Eksperimen dilakukan untuk menemukan model Naïve Bayes yang paling optimal, yaitu dengan menemukan performa terbaiknya. Pada tahap ini akan dilakukan beberapa ekperimen dengan menyelidiki penerapan beberapa langkah text preprocessing yaitu penerapan stopword removal, normalisasi dan stemming. Pada tabel dibawah ini disajikan hasil dari eksperimen setup tersebut.

Tabel 5. Hasil Eksperimen

Ekperimen	Normalisasi	Stopword Removal	Stemming	F1_Score	Accuracy
1	☐	☐	☐	59,64%	62,78%
2	☐	☐	☐	58,77%	61,81%
3	☐	☐	☐	55,06%	56,63%

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 254
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

Ekperimen	Normalisasi	Stopword Removal	Stemming	F1_Score	Accuracy
4	☐	☐	☐	57,35%	60,51%
5	☐	☐	☐	59,70%	63,43%
6	☐	☐	☐	59,01%	62,13%
7	☐	☐	☐	58,37%	60,19%
8	☐	☐	☐	57,11%	59,22%

Berdasarkan eksperimen yang dilakukan, maka ditemukan performa model yang akan di terapkan yaitu dengan mgnggunakan prosos stopword removal dan stemming lebih baik dalam mengklasifikasi kelas pada dataset.. Dengan nilai F1- Score 59,70% dan nilai Accuracy 63,43%. Setelah menemukan performa model yang akan diterapkan langkah selanjutnya atau pengujian model menggunakan data uji.

3.6 Eksperimen Terhadap Jumlah Dataset

Jumlah kelas yang tidak seimbang dapat mempengaruhi tingkat akurasi model dalam mengklasifikasikan data dengan optimal. Oleh karena itu dilakukan pengujian terhadap jumlah data seimbang dan jumlah data tidak seimbang untuk menemukan hasil yang lebih baik dalam mengklasifikasikan data. Berikut ini disajikan dalam bentuk tabel hasil dari kedua eksperimen tersebut.

Tabel 6. Perbandingan Dataset

No	Jumlah Dataset	F1-Score	Accuracy
1	Dataset Seimbang	59,70%	63,43%
2	Dataset tidak Seimbang	41,94%	72,81%

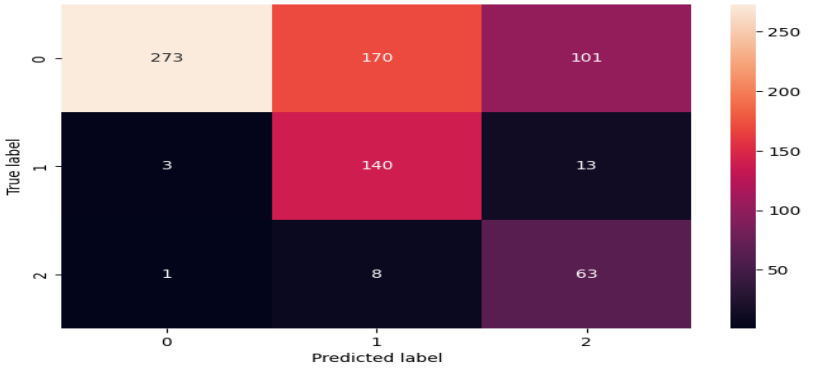
Dari tabel di atas, dapat disimpulkan bahwa pengujian ini menggunakan distribusi data yang merata, dan akurasi model dalam memprediksi kelas dalam kumpulan data jauh lebih tinggi dibandingkan dengan kumpulan data yang tidak seimbang. Hal ini dapat diamati dengan memeriksa data F1-Score yang diperoleh dari kedua pengujian tersebut. Dengan demikian, dalam penelitian ini kuantitas data yang akan dimanfaatkan atau banyaknya data tersebar secara merata.

3.7 Klasifikasi Naïve Bayes

Algoritma Naïve Bayes digunakan untuk tujuan pengujian, dengan rasio pembagian data masing-masing 90%:10% kemudian diambil 10% untuk data pengujian dan validasi. Hasil penelitian menunjukkan nilai F1-Score sebesar 58,72% dan akurasi 61,65%. Hasil ini menunjukkan bahwa metode pengklasifikasi naïve Bayes sangat mahir dalam klasifikasi sentimen.

3.8 Confussion Matrix

Confussion Matrix adalah metrik performa yang digunakan dalam tugas klasifikasi pembelajaran mesin dengan banyak kelas sebagai keluaran. Confussion Matrix adalah representasi tabel yang menampilkan empat kemungkinan kombinasi nilai prediksi dan nilai aktual. Pengujian matriks kebingungan dilakukan pada subset data validasi, yang terdiri dari 10% dari total, untuk menghitung nilai akurasi serta nilai presisi, perolehan, dan Skor F1. Berikut ini disajikan hasil confusion matriks pada penelitian ini.



Gambar 3. Hasil Confusion Matrix

4. KESIMPULAN

Penelitian ini menggunakan metode naïve Bayes classifier untuk mengkategorikan sentimen komentar YouTube seputar pembatalan Indonesia sebagai tuan rumah Piala Dunia U-20. Naïve Bayes Classifier (NBC) adalah pengklasifikasi probabilistik langsung yang memanfaatkan Teorema Bayes dalam kondisi independensi yang kuat. Dataset awal terdiri dari 4629 komentar YouTube. Setelah menerapkan metode pemfilteran, dataset dikurangi

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 255
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 5, Nomor 2, Desember 2023
e-ISSN 2685-998X
DOI 10.30865/json.v5i2.7096

menjadi 3857 komentar. Selanjutnya, dataset tersebut menjalani tahap pelabelan, menghasilkan 2.690 kelas negatif, 386 kasus sentimen positif, dan 386 kasus sentimen netral 781. Pengujian dengan menggunakan dataset yang tidak seimbang menghasilkan tingkat akurasi yang lebih besar yaitu sebesar 72,81%, sedangkan pengujian dengan dataset yang seimbang menghasilkan tingkat akurasi sebesar 61,65%. Namun, untuk memastikan model yang optimal dalam mengklasifikasikan data, keakuratan Skor F1 juga harus diperhitungkan. Nilai F1-Score jauh lebih tinggi yaitu sebesar 59,70% untuk dataset seimbang dibandingkan dengan dataset tidak seimbang yang hanya mencapai nilai F1-Score sebesar 41,94%. Hal ini menunjukkan bahwa model menunjukkan performa unggul dalam melakukan tugas klasifikasi saat menggunakan data seimbang. Penelitian ini menggunakan pendekatan Naïve Bayes yang menghasilkan F1-Score sebesar 58,72% dan tingkat akurasi sebesar 61,65%. Berdasarkan hasil klasifikasi ini, dapat disimpulkan bahwa algoritma Naïve Bayes efektif dalam melakukan klasifikasi sentimen pada komentar YouTube. Temuan akhir menunjukkan bahwa mayoritas masyarakat Indonesia menyatakan sentimen negatif terhadap keputusan Indonesia yang batal menjadi tuan rumah Piala Dunia 2020.

REFERENCES

[1] E. E. Pratama, “Information Retrieval pada Proses Penyimpanan dan Pencarian Dokumen Digital Menggunakan Metode Text Mining Enda Esyudha Pratama,” Fak. Tek. Jur. Inform., pp. 736–742, 2018.
[2] M. A. R. Yulian Findawati, Buku Ajar Text Mining. 2020.
[3] A. V. Sudiantoro et al., “Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier,” Din. Inform., vol. 10, no. 2, pp. 398–401, 2018.
[4] O. W. Purbo, Text Mining. Andi Publlisher, 2019.
[5] F. V. Sari and A. Wibowo, “Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi,” J. SIMETRIS, vol. 10, no. 2, pp. 681–686, 2019.
[6] R. S. N. Yesta Chriptopherus Asia Sanjaya, “‘Kilas Balik Indonesia Ditunjuk Jadi Tuan Rumah Piala Dunia U-20 yang Kini Terancam Batal,’” Kompas.com, 2023. https://www.kompas.com/tren/read/2023/03/27/190000265/kilas-balik-indonesia-ditunjuk-jadi-tuan-rumah-piala-dunia-u-20-yang-kini?page=all
[7] N. S. P. Juana, E. Haerani, F. Syafria, and E. Budianita, “Analisis Sentimen Tanggapan Masyarakat Terhadap Calon Presiden 2024 Ridwan Kamil Menggunakan Metode Naive Bayes Classifier,” J. Sist. Komput. dan Inform., vol. 4, no. 4, p. 570, 2023, doi: 10.30865/json.v4i4.6168.
[8] E. Fitri, “Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes, Random Forest Dan Support Vector Machine,” J. Transform., vol. 18, no. 1, p. 71, 2020, doi: 10.26623/transformatika.v18i1.2317.
[9] Yuyun, Nurul Hidayah, and Supriadi Sahibu, “Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter,” J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 4, pp. 820–826, 2021, doi: 10.29207/resti.v5i4.3146.
[10] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, “Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN,” J. KomtekInfo, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.
[11] Y. Nurtikasari, Syariful Alam, and Teguh Iman Hermanto, “Analisis Sentimen Opini Masyarakat Terhadap Film Pada Platform Twitter Menggunakan Algoritma Naive Bayes,” INSOLOGI J. Sains dan Teknol., vol. 1, no. 4, pp. 411–423, 2022, doi: 10.55123/insologi.v1i4.770.
[12] H. A. R. Harpizon, R. Kurniawan, Iwan Iskandar, R. Salambue, E. Budianita, and F. Syafria, “Analisis Sentimen Komentar Di YouTube Tentang Ceramah Ustadz Abdul Somad Menggunakan Algoritma Naïve Bayes,” JNKTI (Jurnal Nas. Komputasi dan Teknol. Informasi), vol. 5, no. 1, pp. 131–140, 2022, [Online]. Available: http://repository.uin-suska.ac.id/59746/
[13] M. R. Amly, Y. Yusra, and M. Fikry, “Penerapan Metode Naïve Bayes Classifier Pada Klasifikasi Sentimen Terhadap Anies Baswedan Sebagai Bakal Calon Presiden 2024,” J. Sist. Komput. dan Inform., vol. 4, no. 4, p. 621, 2023, doi: 10.30865/json.v4i4.6214.
[14] K. V. S. Toy, Y. A. Sari, and I. Cholissodin, “Analisis Sentimen Twitter menggunakan Metode Naive Bayes dengan Relevance Frequency Feature Selection (Studi Kasus: Opini Masyarakat mengenai Kebijakan New Normal),” J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 11, pp. 5068–5074, 2021, [Online]. Available: http://j-ptiik.ub.ac.id
[15] Vynska Amalia Permadi, “Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Review Restoran di Singapura,” J. Buana Inform., vol. 11, pp. 141–151, 2020.
[16] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, “Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile,” Petir, vol. 15, no. 2, pp. 264–275, 2022, doi: 10.33322/petir.v15i2.1733.
[17] R. Ghaniy and K. Sihotang, “Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir,” Teknois J. Ilm. Teknol. Inf. dan Sains, vol. 9, no. 1, pp. 63–72, 2019, doi: 10.36350/jbs.v9i1.7.
[18] S. Chohan, A. Nugroho, A. M. B. Aji, and W. Gata, “Analisis Sentimen Pengguna Aplikasi Duolingo Menggunakan Metode Naïve Bayes dan Synthetic Minority Over Sampling Technique,” Paradig. - J. Komput. dan Inform., vol. 22, no. 2, pp. 139–144, 2020, doi: 10.31294/p.v22i2.8251.
[19] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, “Dataset Indonesia untuk Analisis Sentimen,” J. Nas. Tek. Elektro dan Teknol. Inf., vol. 8, no. 4, p. 334, 2019, doi: 10.22146/jnteti.v8i4.533.
[20] L. Ardiani, H. Sujaini, and T. Tursina, “Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak,” J. Sist. dan Teknol. Inf., vol. 8, no. 2, p. 183, 2020, doi: 10.26418/justin.v8i2.36776.
[21] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, “Pengaruh Text Preprocessing terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19),” J. Media Inform. Budidarma, vol. 5, no. 2, p. 406, 2021, doi: 10.30865/mib.v5i2.2835.
[22] A. P. Giovani, A. Ardiansyah, T. Haryanti, L. Kurniawati, and W. Gata, “Analisis Sentimen Aplikasi Ruang Guru Di

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 256
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Twitter Menggunakan Algoritma Klasifikasi," J. Teknoinfo, vol. 14, no. 2, p. 115, 2020, doi: 10.33365/jti.v14i2.679.
[23] G. A. Mursianto, D. Widiyanto, and B. T. Wahyono, "Analisis Sentimen Ulasan Pengguna Pada Aplikasi Google Classroom Menggunakan Metode SVM Dan Seleksi Fitur PSO," Inform. J. Ilmu Komput., vol. 18, no. 3, p. 221, 2022, doi: 10.52958/iftk.v18i3.4685.
[24] N. L. P. C. Savitri, R. A. Rahman, R. Venyutzky, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning," J. Tek. Inform. dan Sist. Inf., vol. 7, no. 1, pp. 47–58, 2021, doi: 10.28932/jutisi.v7i1.3216.

Copyright © 2023 **Ilham Habibi Hasibuan**, Page 257
This Journal is licensed under a Creative Commons Attribution 4.0 International License