Analisis Sentimen Pengguna Media Sosial Twitter Tentang Gelaran Piala Asia Qatar dengan Metode Naive Bayes

Aziz Musthafa[1], Dihin Muriyatmoko[2], Ahmad Fa'iq Dzulfikar As'ad [3*]
[123]Universitas Darussalam Gontor
[123]Ponorogo, Jawa Timur-Indonesia
Email : [1] aziz@unida.gontor.ac.id , [2] dihin@unida.gontor.ac.id ,
[3]ahmadasad42007@mhs.unida.gontor.ac.id

Abstract

In the digital era and technological advancement, social media has become an important platform for users to share thoughts, opinions and feelings on various topics, including sporting events such as the Asian Cup. Therefore, it is important to understand people's views and opinions by analyzing social media data. This analysis classifies everything into three categories: positive, negative and neutral based on predetermined classes. This data analysis technique uses CRISP-DM, an industry standard data mining process, starting from business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Features are selected using the Query Expansion Ranking technique so that all data is collected based on certain classes. The number of features required to improve accuracy. The next stage is the application of the classification algorithm technique, namely the Naive Bayes technique. The classification results using the Naive Bayes method in this study had an accuracy rate of 91%. After the validation process with K-fold cross validation, the accuracy value obtained from Naïve Bayes was 91%. Based on the model classification results, neutral sentiment dominates with an accuracy result of 64.1%, for positive sentiment the accuracy result is 33.6% and for negative sentiment the accuracy result is 2.3%. These accuracy results show that a lot of data contains neutrality towards the 2024 Asian Cup in Qatar.

Keywords: *Sentiment Analysis, Asian Cup, Social Media X, Naive Bayes.*

Abstraksi

Di era digital dan kemajuan teknologi, media sosial telah menjadi platform penting bagi pengguna untuk berbagi pemikiran, pendapat, dan perasaan tentang berbagai topik, termasuk acara olahraga seperti Piala Asia. Oleh karena itu, penting untuk memahami pandangan dan opini masyarakat dengan menganalisis data media sosial. Analisis ini mengklasifikasikan keseluruhan menjadi tiga kategori: positif, negatif, dan netral berdasarkan kelas yang telah ditentukan. Teknik analisis data ini menggunakan CRISP-DM, sebuah proses penambangan data standar industri, dimulai dari businesss understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Fitur dipilih menggunakan teknik Query Expansion Ranking sehingga semua data dikumpulkan berdasarkan kelas tertentu. Jumlah fitur yang dibutuhkan untuk meningkatkan akurasi. Tahap selanjutnya yaitu penerapan teknik algoritma klasifikasi yaitu teknik Naive Bayes. Hasil klasifikasi menggunakan metode Naive Bayes pada penelitian ini memiliki tingkat akurasi sebesar 91%. Setelah proses validasi dengan K-fold cross validation, nilai akurasi

yang didapat pada Naïve Bayes adalah 91%. Berdasarkan hasil klasifikasi model, sentimen netral mendominasi dengan hasil akurasi sebesar 64.1%, untuk sentimen positif memiliki hasil akurasi sebesar 33,6% dan pada sentimen negatif hasil akurasi yang dimiliki sebesar 2,3%. Dari hasil akurasi tersebut menunjukkan bahwa banyak data yang berisi kenetralan terhadap gelaran Piala Asia 2024 di Qatar.

Kata Kunci: *Analisis Sentimen, Piala Asia, Media Sosial X, Naive Bayes*

1. PENDAHULUAN

Di era digital dan kemajuan teknologi, media sosial telah menjadi platform penting bagi pengguna untuk berbagi pemikiran, pendapat, dan perasaan tentang berbagai topik, termasuk acara olahraga seperti Piala Asia. Oleh karena itu, penting untuk memahami pandangan dan opini masyarakat dengan menganalisis data media sosial.

Piala Asia adalah turnamen sepak bola yang diselenggarakan oleh *Asian Football Confederation* (AFC). Piala Asia 2024 Qatar mendapatkan giliran sebagai tuan rumah pelaksanaan Piala Asia edisi ke 18. Acara tersebut secara resmi berlangsung pada musim dingin dan satu bulan lamanya, yaitu sejak tanggal 12 Januari hingga 10 Februari 2024 dengan partisipan sebanyak 24 Tim. Twitter merupakan salah satu media koneksi yang dicari semua orang di seluruh dunia. Hal ini dibuktikan dengan semakin meningkatnya jumlah pengguna Twitter di seluruh dunia [1], termasuk Indonesia. Saat ini, Twitter memiliki 313 juta pengguna aktif bulanan pada tahun 2016 [2]. Pengguna Twitter memberikan berita dan komentar terkini tentang topik paling penting di dunia saat ini. Topik-topik besar yang sedang hangat dan sering dikomentari pengguna menimbulkan isu dan polemik di media sosial khususnya Twitter. Oleh karena itu, reaksi atau sentimen masyarakat terhadap Piala Asia Qatar 2024 positif, negatif, atau netral. Untuk melihat reaksi tersebut, peneliti menggunakan salah satu platform media sosial, Twitter, untuk memperoleh data. Data yang dihasilkan oleh Twitter, akan diproses dan dianalisis dengan benar, dapat menjadi penting dan berguna bagi masyarakat dan organisasi [3]**.**

Analisis sentimen termasuk dalam bidang *Natural Language Processing* (NLP) yang mengidentifikasi isi data berupa teks positif, negatif, atau netral berupa pendapat dan pandangan (sentimen) tentang suatu topik atau peristiwa tertentu [4,5]. Analisis sentimen pada suatu kalimat menggambarkan bagian pertimbangan penilaian terhadap entitas atau kejadian tertentu [6]. Berdasarkan pemaparan masalah, maka pada penelitian ini akan dilakukan analisis sentimen pengguna media sosial Twitter terhadap Piala Asia 2024 di Qatar tahun 2024 menggunakan algoritma *Naïve Bayes*. Dengan bertujuan untuk

mengetahui sentimen dan hasil ketepatan klasifikasi sentimen pengguna media sosial Twitter tentang gelaran Piala Asia di Qatar pada tahun 2024 menggunakan *Naïve Bayes*. Proses analisis ini dilakukan dengan menggunakan pendekatan *Text Mining*, dan *Machine Learning* dalam pelaksanaannya hingga menghasilkan kesimpulan opini masyarakat mengenai gelaran acara tersebut dan juga menguji tingkat akurasi, presisi, recall, dan f1-score pada metode *Naive Bayes Classifier* dengan TF Idf dalam mengklasifikasikan data tweet tentang gelaran Piala Asia 2024 [7]. Analisis sentimen atau *opinion mining* merupakan salah satu bidang data mining yang biasa digunakan untuk menganalisis data teks berupa opini yang terpolarisasi sehingga diperoleh informasi yang bernilai positif, negatif, atau netral. [8].

Naive Bayes adalah metode probabilitas yang pertama kali diperkenalkan oleh ilmuwan Inggris Thomas Bayes. *Naive Bayes* digunakan untuk memprediksi peluang masa depan berdasarkan pengalaman masa lalu [9,10]. Menerapkan *Naive Bayes* pada analisis sentimen memerlukan dua proses penting: pelatihan dan pengujian. Keuntungan *Naive Bayes* adalah menggunakan lebih sedikit data pelatihan, sehingga penghitungan dapat dilakukan lebih cepat dan efisien. Kelemahan *Naive Bayes* adalah pemilihan fitur yang salah mengurangi akurasi dan menambah waktu komputasi.

Berikut persamaan dari teorema Bayes [11] :

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (1)$$

dimana :

B : *Class* yang belum diketahui

A : Hipotesis data B

(A|B) : Probabilitas hipotesis A berdasarkan kondisi B (*posterior probability*)

(A) : Probabilitas hipotesis A (*prior probability*)

(B|A) : Probabilitas B berdasarkan kondisi pada hipotesis A

(B) : Probabilitas B

2. METODE PENELITIAN

Analisis ini mengklasifikasikan keseluruhan menjadi tiga kategori: positif, negatif, dan netral berdasarkan kelas yang telah ditentukan. Teknik analisis data ini menggunakan CRISP-DM, sebuah proses penambangan data standar industri. CRISP-DM seperti terlihat pada Gambar 1, dimulai dari *businees understanding*, *data understanding*, *data*

preparation, *modelin*g, *evaluation*, dan *deployment*. Fitur dipilih menggunakan teknik *Query Expansion Ranking* sehingga semua data dikumpulkan berdasarkan kelas tertentu. Jumlah fitur yang dibutuhkan untuk meningkatkan akurasi. Tahap selanjutnya yaitu penerapan teknik algoritma klasifikasi yaitu teknik *Naive Bayes*.

Gambar 1. Metode Penelitian CRISP-DM

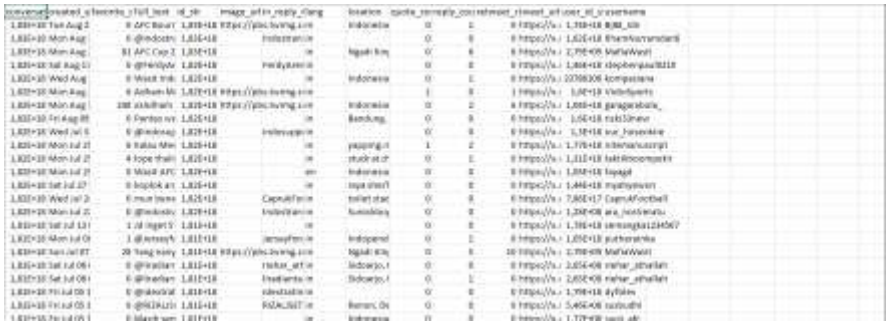
2.1. Business Understanding

Ini adalah tahap pertama dalam melakukan CRISP-DM. Tahapan ini dilaksanakan penyatuan ide penulis atas permasalahan yang sedang populer di lingkungan masyarakat. Berdasarkan rumusan masalah, bahwa permasalahan yang diangkat dalam karya tulis ini adalah banyaknya opini masyarakat yang diungkapkan pada media sosial Twitter. Khususnya tanggapan terhadap Piala Asia 2024. Ide ini di angkat berdasarkan permasalahan yang masih menjadi polemik di masyarakat.

2.2. Data Understanding

Pada tahap ini penulis melakukan proses pengambilan data dan pelabelan data, proses pengambilan data dengan cara *crawling* data. Data diambil secara acak sampai dengan 5224 data terkumpul. Dengan kata kunci dibagi menjadi 3 yaitu “Naturalisasi”, “AsianCup2023”, “Wasit AFC”.

Kemudian Pelabelan data dilakukan secara manual pada 2050 data berdasarkan tiga kategori sentimen yaitu positif, negatif dan netral. Dengan jumlah data 2050 dan data ini memiliki Atribut yang digunakan dalam penelitian ini meliputi: `conversation_id_str`, `created_at`, `favorite_count`, `full_text`, `id_str`, `image_url`, `in_reply_id`, `lang`, `location`, `quote_count`, `reply_count`, `retweet_count`, `tweet_url`, `user_id_str`, dan `username`. Kelas yang dianalisis adalah: positif, negatif, netral. Dalam penelitian ini, sentimen yang dianalisis berfokus pada atribut `full_text`. Data kotor yang telah dikumpulkan dapat dilihat pada Gambar 2.



Gambar 2. Proses pengambilan data menggunakan teknik *crawling*

2.3. Data Preparation

Selanjutnya dilakukan *Data Preparation*. Data yang akan dimasukkan dalam kategori Positif adalah data *tweet* yang berisi pernyataan mendukung Piala Asia 2024. Data Netral adalah data *tweet* yang hanya berisi informasi atau berita. Data Negatif adalah data *tweet* yang menunjukkan bahwa mereka tidak mendukung jalannya Piala Asia 2024. Sebelum data diklasifikasi, harus melalui *pre-processing* seperti *case folding*, *tokenizing*, *filtering*, dan *stemming* agar dapat diurutkan dan dikelompokkan secara sistematis. Berikut adalah tahapan dari data *preprocessing*:

2.3.1 Data Preprocessing

Pada tahap ini akan dilakukan langkah-langkah untuk membersihkan data teks lalu data teks dipersiapkan untuk pengolahan nantinya. Langkah data *preprecossing* diperlukan untuk menyelesaikan beberapa jenis masalah termasuk *noisy data*, data redundansi, nilai data yang hilang, dan lain lain [10]. Adapun langkah-langkah data *preprocessing* adalah *case folding, tokenizing, filtering, stemming*.

- Langkah pertama adalah *Case Folding,* Selama fase ini, semua karakter dalam dokumen atau kalimat diubah menjadi huruf kecil [12].
- Langkah kedua adalah *Tokenizing,* pada tahap ini merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu [13].
- Langkah ketiga adalah *Filtering*, tahap di mana kata-kata penting diambil dari hasil term. Contoh *stopwords* termasuk "yang", "dan", "di", "dari", dan sebagainya.
- Langkah terakhir adalah *Stemming, Stemming* digunakan untuk menyegeramkan kata dan memperkecil daftar kata pada data latih [14].

2.3.2 Pembobotan TF-IDF

Pada tahap ini akan diteliti seberapa sering suatu kata muncul dalam suatu dokumen. Metode yang digunakan untuk mengetahuinya adalah dengan pembobotan TFIDF (*Term Frequency - Invers Document Frequency*).

Tujuan dari TF-IDF ada 2 yaitu yaitu:

1. TF (*Term Frequency*): menghitung seberapa sering suatu kata/token yang muncul dalam dokumen. Rumus TF sesuai dengan persamaan 1 berikut ini:

$$TF(tn, dn) = f(tn, dn) \quad (2)$$

   Jadi $f(tn, dn)$ mendefinisikan jumlah kemunculan term-*n* pada sebuah dokumen-*n* [15].

2. IDF (*Invers Document Frequency*): menilai seberapa penting suatu kata/token. Rumus dari IDF sesuai dengan persamaan 2 berikut ini:

$$IDF\ (tn) = \log \frac{D}{df(t)} \quad (3)$$

Jadi $\log \frac{D}{df(t)}$ Menjelaskan jumlah dokumen pada dataset-*D* akan dibagi dengan jumlah dokumen yang mengandung term-df(t)

Lalu menghitung keseluruhan nilai TF-IDF dengan mengkalikan nilai keduanya. Rumus sesuai dengan persamaan 3 berikut:

$$Wtd = TF(tn, dn) \times IDF\ (tn) \quad (4)$$

Keterangan:

W : Bobot kata-t terhadap dokumen ke-d.

t : Term atau kata ke-n dari kata yang di cari.

d : Dokumen ke-n.

TF : Frekuensi/jumlah Sebuah kata dalam satu dokumen.

IDF : *Invers Documen Frequency.*

D : Total Dokumen.

Df : Jumlah dokumen yang mengandung Term/Kata yang di cari [16] .

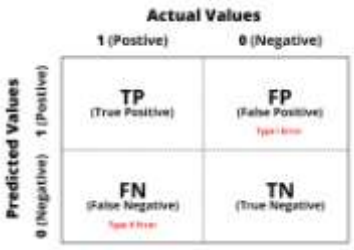
2.4. Modelling

Pada tahap ini data dibagi menjadi data *training* dan data *testing*, data *training* merupakan data yang digunakan untuk mempelajari pola dan karakteristik data, sedangkan data *testing* merupakan data yang digunakan untuk menguji model yang telah

mempelajari pola dan karakteristik data. Data *training* dan data *testing* ini mendapatkan 70% data *training* dan juga 92% data *testing*. Metode yang digunakan untuk klasifikasi ini adalah metode klasifikasi *Naïve Bayes*.

2.5. Evaluation

Pada tahap pengolahan ini dilakukan pembelajaran data untuk mengenali pola dari data teks yang diolah dan diberi bobot. Setelah model mengenali pola di setiap kelas, model dapat mengklasifikasikan data baru. Kemudian setelah data diuji dan diolah maka akan dilakukan pengujian terhadap performa dan efektifitas dari sistem menggunakan *K-fold cross validation* dan menggunakan *Confusion Matrix*. *K-fold Cross Validation* adalah teknik yang umum digunakan dalam pembelajaran mesin dan penambangan teks untuk mengevaluasi kinerja model. Ini melibatkan pembagian data kedalam himpunan bagian, pelatihan dan pengujian berulang-ulang himpunan bagian sehingga setiap titik data diambil sampelnya dan dilatih. Pada Gambar 3 menunjukkan komponen utama dari *Confusion Matrix*.



Gambar 3. Confusion Matrix

True Positives (TP): memprediksi data positif dengan benar.
True Negatives (TN): memprediksi data positif dengan benar.
False Positive (FP): Data negatif namun diprediksi sebagai data positif.
False Negative (FN): Data positif tetapi diprediksi sebagai data negatif.

2.6. Deployment

Dalam tahapan hasil dari sistem akan digunakan oleh masyarakat secara luas dan lembaga yang berperan dalam mengembangkan sepak bola terkhusus kepada Tim Nasional Indonesia agar mempermudah dalam klasifikasi berita yang tersebar di masyarakat agar bisa merespon berbagai isu yang tersebar selama acara berlangsung ataupun sudah selesai dan dijadikan sebagai acuan untuk bijak dalam berkomentar.

3. HASIL DAN PEMBAHASAN

Pada tahapan ini berisi pembahasan dan hasil dari proses implementasi yang selesai dilakukan. Penelitian ini memiliki tujuan untuk memberi klasifikasi sentimen masyarakat pada media sosial Twitter mengenai pelaksanaan Piala Asia 2024 di Qatar menggunakan metode algoritma *Naïve Bayes*. Data dikumpulkan menggunakan teknik crawl pada *notebook google colab* dengan data yang bersumber dari media sosial Twitter. Data yang terkumpul dibagi menjadi 3 kategori, yaitu positif, negatif dan netral.

3.1. Pengumpulan Data

Pada penelitian ini pengumpulan data dilakukan pada *google collab* dengan memanfaatkan *library tweet harvest* dengan Bahasa *python*, data yang digunakan diambil dari media sosial Twitter dengan kata kunci "Naturalisasi", "Wasit AFC", dan "*AsianCup*2023". Data diambil secara acak dari tanggal 1 November 2023 hingga 1 Maret 2024 sebanyak 5324 data. Setelah melalui proses pembersihan secara manual dengan membuang tweet yang duplikat menggunakan Microsoft Excel didapati 2050 data pelatihan model dan 1314 baru untuk menguji model yang telah dilatih. Contoh data yang berhasil dikumpulkan dapat dilihat pada Tabel 4.1.

Tabel 1. Data yang terkumpul

No	Date	Komentar	Username
1	Fri Feb 02 09:55:14 +0000 2024	Aymen Hussein dikartu merah usai merayakan gol untuk Irak di babak 16 besar Piala Asia 2023. AFC menegaskan keputusan wasit sudah tepat.	detiksport
2	Sat Jan 13 13:57:11 +0000 2024	Wasit asal Jepang Yoshimi Yamashita menjadi wasit putri pertama yang memimpin pertandingan AFC Asian Cup. *#AsianCup2023*	sporttiaphari
3	Mon Jan 15 11:17:05 +0000 2024	@Indostransfer Fariq Hitaba harusnya ikut jadi wasit AFC Cup	febriana_riady

3.2. Pelabelan Data

Sebelum masuk ke tahap *processing*, data terlebih dahulu dilakukan cara *preprocessing*. Sentimen dalam data tersebut dibagi menjadi 3 label yaitu positif, negatif, dan netral. Data yang sudah terkumpul dan diberi label akan dibersihkan dengan tahapan-tahapan *preprocessing* diantaranya *case folding*, *tokenizing*, *cleaning* dan *stemming*. Pada proses ini data dilabeli melalui 3 tahapan, pertama dilabeli secara individu kemudian kedua melalui dosen ahli dan kemudian ketiga melalui *google colab notebook*. Pada tahapan melalui dosen ahli, penulis memberikan data kepada Ustadzah Nurhanna

Marantika M.A untuk diverifikasi lebih detail pada data yang telah terkumpul. Proses ini dilakukan dengan memasukkan data yang sudah dikumpulkan dan diberi label sebelumnya. Contoh hasil dari proses pelabelan data bisa dilihat pada Tabel 2.

Tabel 2. Pelabelan Data

No	Komentar	Label
1	Wasit problematik dinegaranya dipakai AFC. malu2in AFC banyak di weibo skandal nih wasit	Netral
2	Dari banyaknya pemain naturalisasi kayaknya dia doang yg paling star syndrom. Yg lain biasa² aja. Elkan yg paling pertama dateng paling kalem. Nathan yg paling banyak diidolain juga kalem² aja	Positif
3	Kena Penalti emang masalah dari dulu piala AFC kemaren kena Penalti juga. Mending evaluasi mencegah nya daripada nyalahin wasit.. timnas senior pas kualifikasi kemaren aja bisa main bersih.. Emang tergantung kualitas pemain aja sih ini	Positif

Proses pelabelan kali ini membutuhkan waktu selama 1 Bulan, dalam proses pelabelan juga melakukan pembersihan data secara manual. Dari ketiga kategori yang paling mendominasi adalah sentimen netral dengan jumlah data sebanyak 1314, kemudian disusul sentimen positif sebanyak 689 data kemudian sentimen negatif dengan jumlah 47 data. Untuk hasil perbandingan ketiga kategori dapat dilihat pada Tabel 3.

Tabel 3. Data yang terkumpul

No	Kelas	Jumlah
1	Positif	689
2	Negatif	47
3	Netral	1314
Total		2050

3.3. Preprocessing

Pada proses ini dilakukan pemrosesan pada data teks untuk dibersihkan dan diseragamkan sehingga dapat memaksimalkan klasifikasi program kepada hasil yang diinginkan. pada proses ini dilakukan beberapa pemrosesan pada data teks yang sebagai berikut :

(1) Case Folding

Pada proses ini data teks diseragamkan menjadi huruf kecil dan dibersihkan dengan beberapa pengaturan perubahan yang dibutuhkan untuk memaksimalkan proses klasifikasi dari program. Perubahan yang dilakukan adalah mengubah isi data menjadi huruf kecil, menghapus *hyperlink*, menghapus tanda koma, menghapus angka,

menghapus semua karakter yang bukan huruf dan spasi. Tabel 4 menunjukan hasil perubahan pada data mentah hingga perubahan setelah dilakukan *Case Folding*.

Tabel 4. Hasil *Case Folding*

Data Awal	*Case Folding*
Bayangin kamu lagi gak enak badan. Terus Boomer Boomer brengsek mau naturalisasi dokter. Elon yang bukan pedagang molen di jatihandap nanam starlink tai di puskesmas Bali. Dan inpluenser udah dimainkan	bayangin kamu lagi gak enak badan. terus boomer boomer brengsek mau naturalisasi dokter. elon yang bukan pedagang molen di jatihandap nanam starlink tai di puskesmas bali. dan inpluenser udah dimainkan.

(2) Tokenizing

Tahapan *preprocessing* selanjutnya adalah *tokenizing*, setiap teks akan dibagi menjadi bagian yang lebih kecil, bagian-bagian tersebut umumnya disebut token. Tujuan *tokenizing* adalah untuk memecah teks menjadi beberapa bagian yang lebih mudah diproses untuk analisis. Hasil dari *tokenizing* dapat dilihat pada Tabel 5.

Tabel 5. Hasil *Tokenizing*

Data Awal	*Tokenizing*
Lah bentar lagi naturalisasi warga Singapore ke ikn loh...mantap kan	['lah', 'bentar', 'lagi', 'naturalisasi', 'warga', 'singapore', 'ke', 'ikn', 'loh', 'mantap', 'kan']

(3) Filtering

Pada tahapan ini data teks disaring dari kata-kata yang kurang berguna pada proses klasifikasi program menggunakan *library stopword* bahasa indonesia yang ada pada *library* NLTK (*Natural Language Tool Kit*). Setelah proses *filtering* data yang kurang berguna maka didapatkan hasil dari *filtering* pada Tabel 6 di bawah ini.

Tabel 6. Hasil *Filtering*

Data Awal	*Filtering*
Tetap optimis untuk Timnas Indonesia U-23 Buruan Intip! Link Live Streaming Indonesia vs Uzbekistan di Semifinal Piala Asia U-23 2024 #TimnasDay #KitaGaruda #BersamaGaruda #GarudaMendunia #AsianCup2023 #PialaAsia	tetap optimis untuk timnas indonesia u buruan intip link live streaming indonesia vs uzbekistan di semifinal piala asia u timnasday kitagaruda bersamagaruda garudamendunia asiancup pialaasia

(4) Stemming

Tahapan Selanjutnya atau tahapan terakhir dalam *preprocessing* adalah *Stemming*. Dalam tahapan *stemming* ini kata – kata dalam teks diubah ke kata dasarnya atau akarnya dengan menghilangkan akhiran atau awalan tertentu. Tujuan *stemming* adalah untuk

mengurangi variasi dalam kata atau entitas yang sama. Hasil proses *stemming* ditunjukkan pada Tabel 7.

Tabel 7. Hasil *Stemming*

Data Awal	*Stemming*
Paling utama naturalisasi pejabat dan menteri krn banyak kebijakan yg di luar nalar masyarakat awam.	utama naturalisasi jabat menteri krn bijak yg nalar masyarakat awam

3.4. Pembobotan TF-IDF

Data yang melewati proses *preprocessing* diberi bobot menggunakan teknik pembobotan TF-IDF. Teknik ini digunakan dalam pemrosesan teks untuk menilai seberapa penting sebuah kata dalam suatu dataset. Teknik ini Merupakan teknik yang umum untuk pemrosesan kata dan penambangan data. Manfaat dalam penggunaan pembobotan ini dapat membantu mengatasi masalah ketidakseimbangan kata kemudian diberikan Pembobotan kata yang lebih akurat dan berdampak besar pada analisis sentimen.

Gambar 4. Hasil Pembobotan TF-IDF

Pada gambar 4 diatas menunjukan hasil dari Pemrosesan TF-IDF dari berbagai kata yang ada di dalam dataset. Angka pertama dalam setiap tuple menunjukkan indeks dokumen. Angka kedua dalam setiap tuple menunjukkan indeks fitur (kata) dalam matriks fitur. Angka ketiga menunjukkan hasil dari TF-IDF dari kata tersebut dalam dokumen.
Contoh baris pertama dari gambar: (0, 383) 0.16414040239207908 artinya:

- Kata dengan indeks 383 muncul dalam dokumen dengan indeks 0.
- Nilai TF-IDF dari kata tersebut dalam dokumen ini adalah 0.1641.

3.5. Evaluation

3.5.1. Hasil Evaluasi Confusion Matrix Naive Bayes

Setelah tahapan *Preprocessing* dan pembobotan menggunakan TF-IDF, pemodelan dilakukan dengan algoritma. Model yang diterapkan adalah Metode *Naïve Bayes*.

Pengklasifikasian ini menggunakan data *training* sebanyak 2050 data dan data *testing* sebanyak 1314 data. Data yang sudah melalui tahap pembobotan akan siap untuk melalui tahap proses klasifikasi, data yang sudah diproses pada tahap klasifikasi selanjutnya akan melalui tahap evaluasi, evaluasi dilakukan menggunakan *Confusion Matrix*. Hasilnya menunjukkan bahwa *Naive Bayes* berhasil mengklasifikasi data dengan nilai *True Negative* sebesar 242, *True Neutral* sebesar 235, dan *True Positive* sebesar 241. Dan juga *False Negative* sebesar 14, *False Neutral* sebesar 26, pada kelas *False Positive* sebesar 31. Dengan memiliki akurasi rata - rata 0.91.

3.5.2. Hasil Evaluasi Cross Validation Naive Bayes

Tahap selanjutnya adalah validasi, yang dalam penelitian ini dilakukan menggunakan *K-fold Cross Validation*. Sebelum proses validasi dengan *K-fold Cross Validation*, akurasi yang diperoleh adalah 0,91. Setelah melalui tahap validasi menggunakan *K-fold Cross Validation* nilai akurasi rata-rata adalah 0,91, dengan nilai tertinggi mencapai 0,93.

3.6. Deployment

3.6.1. Visualisasi WordCloud

Data yang telah melalui proses *preprocessing* hingga proses evaluasi direpresentasikan dengan menyoroti seberapa pentingnya kata – kata itu dalam data. Kata – kata yang sering muncul akan tampak lebih pekat warnanya dari lain nya. Hasil dari representasi Wordcloud kata yang sering muncul ditunjukan pada Gambar 7.

Gambar 7. Visualisasi *Word Cloud*

3.6.2. Hasil Distribusi Sentimen

Dari hasil klasifikasi model *Naive Bayes*, sentimen netral memiliki persentase akurasi sebesar 64% disusul oleh positif dengan persentase 33% kemudian sentimen

negatif dengan persentase sebesar 2%. Persentase dalam bentuk diagram dapat dilihat pada gambar 8.

Gambar 8. Diagram Hasil Distribusi Sentimen

Dengan jumlah sentimen netral yang tinggi dengan hasil klasifikasi 64% mengindikasikan dimana terdapat banyak data tweet yang menunjukkan kenetralan terhadap penyelenggaraan Piala Asia 2024 di Qatar. Karena pada penyelenggaraan kali ini para peserta perlombaan tidak banyak dirugikan oleh panitia penyelenggara ataupun wasit dan lain sebagainya. Kemudian para penggemar juga antusias mengikuti gelaran Piala Asia Qatar 2024 hingga berakhirnya acara tersebut karena tidak ada nya indikasi kecurangan dan merugikan pihak tim tamu dalam gelaran tersebut. Hasil klasifikasi 33% pada sentimen positif menunjukkan bahwa data tweet yang mengandung sentimen positif juga cukup signifikan menyusul dukungan – dukungan yang ada terhadap timnas Indonesia yang sudah berpartisipasi kembali pada Piala Asia edisi tahun 2024 beserta skuad terbaiknya dan mampu mencapai hingga babak 16 besar. Kemudian untuk sentimen negatif dengan hasil klasifikasi 2%. Hal ini mengindikasikan bahwa penyelenggara Piala Asia tahun ini memberikan dampak yang positif kepada seluruh negara yang mengikuti ajang kali ini, dan perangkat pertandingan yang telah bekerja dengan baik selama jalannya kompetisi ini.

4. KESIMPULAN

Pada penelitian ini dapat disimpulkan sebagai berikut:

1. Model yang dibuat dapat mengklasifikasikan sentimen masyarakat yang diungkapkan di media sosial X (Twitter), khususnya tentang Piala Asia 2024 yang berlangsung pada

tanggal 12 Januari 2024 – 14 Februari 2024 pada kumpulan tweets dalam opini setuju, tidak setuju dan netral menggunakan metode *Naive Bayes.*

2. Klasifikasi yang menggunakan Metode *Naive Bayes* pada penelitian ini memiliki tingkat akurasi sebesar 0,91.
3. Setelah proses validasi dengan *K-fold cross validation*, nilai akurasi yang didapat pada *Naïve Bayes* adalah 0,91.
4. Berdasarkan hasil klasifikasi model, sentimen netral mendominasi dengan hasil klasifikasi sebesar 64.1%, untuk sentimen positif memiliki hasil klasifikasi sebesar 33,6% dan pada sentimen negatif hasil klasifikasi yang dimiliki sebesar 2,3%. Dari hasil persentase tersebut menunjukkan bahwa banyak data yang berisi kenetralan terhadap pegelaran Piala Asia 2024 di Qatar.

5. SARAN

Dengan hasil yang didapat dari penelitian ini, peneliti menyadari bahwa masih banyaknya kekurangan dari penelitian ini, dengan begitu perlunya pengembangan lebih lanjut untuk mendapatkan hasil yang lebih maksimal, berikut beberapa saran yang dapat diambil yaitu:

- Agar menambahkan varian data yang lebih banyak dari berbagai media sosial yang berbeda,
- Memodifikasi pada tahapan *Preprocessing* supaya nantinya dapat memberikan peningkatan akurasi pada klasifikasi,
- Mencoba untuk menambahkan model guna menghasilkan nilai akurasi yang berbeda, kemudian membandingkan nilai dari setiap model.

DAFTAR PUSTAKA

[1] Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA," J. Ilm. Edutic, vol. 7, no. 1, pp. 1–11, 2020.

[2] A. Samad, H. Basari, B.Hussin, I. G. Pramudya, and J. Zeniarja, "*Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization*," *Procedia Eng*., vol. 53, no. *December*, pp. 453–462, 2013, doi: 10.1016/j.proeng.2013.02.059.

[3] Saputra, A., Subing, M., & Pratama, R. (2023). Perbandingan Metode *Naïve Bayes Classifier* Dan *Support Vector Machine* Untuk Analisis Sentimen Pengguna Twitter Mengenai Piala Dunia Fifa 2022. *Teknomatika*, *13*(01), 22–31.

[4] Fanissa, S., Fauzi, A. M. and Adinugroho, S. (2018). "Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode *Naive Bayes* dan Seleksi Fitur *Query Expansion Ranking* | Jurnal Pengembangan Teknologi Informasi, SMATIKA Jurnal Volume 10 Nomor 02, Desember Tahun 2020 ISSN: 2087-0256, e-ISSN: 2580-6939 dan Ilmu Komputer," J. Pengembangan Teknologi Informasi dan Ilmu Komputer., vol. 2, no. 8, pp. 2766– 2770.

[5] Chandani, V., Komputer, I. F. and Nuswantoro, D. U. (2015). "Komparasi Algoritma Klasifikasi *Machine Learning* Dan *Feature Selection* pada Analisis Sentimen Review Film," J. Intell. Syst., vol. 1, no. 1, pp. 56–60.

[6] Prasetya, E., 2006, *Case Based Reasoning* untuk mengidentifikasi kerusakan bangunan, *Tesis,* Program Pasca Sarjana IlmuKomputer, Univ. GadjahMada, Yogyakarta.

[7] Marga, N. S., Isnain, A. R., & Alita, D. (2020). *TERHADAP KASUS CORONA MENGGUNAKAN METODE NAIVE BAYES*. *2*(4), 453–463.

[8] Laurensz, B., & Sediyono, E. (2021). Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (*Analysis of Public Sentiment on Vaccination in Efforts to Overcome the Covid-19 Pandemic*). Jurnal Nasional Teknik Elektro Dan Teknologi Informasi, 10(2), 118–123.

[9] Tuhuteru, H., & Iriani, A. (2018). Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode *Support Vector Machine* dan *Naive Bayes Classifier*. Jurnal Informatika: Jurnal Pengembangan IT, 3(3), 394–401. https://doi.org/10.30591/jpit.v3i3.977

[10] Susanti, D.N., Sediyono, E., dan Sembiring, I. (2016). "Uji Perbandingan Akurasi Analisis Sentimen Pariwisata Menggunakan Algoritma *Support Vector Machine* dan *Naive Bayes*" Nusantara of Engineering, Vol. 3, No. 2, hal. 26–33.

[11] Bustami, "Penerapan Algoritma Naive Bayes," J. Inform., vol. 8, no. 1, pp. 884–898, 2014.

[12] Valatehan, Lucky, Muhammad Fachrurrozi, dan Osvari Arsalan. 2016. Identifikasi Kalimat Pemborosan Menggunakan *Rule Based Reasoning*. *Annual Research* Seminar Vol 2 No. 1. Palembang: Universitas Sriwijaya.

[13] Amin, Fatkhul. 2012. Sistem Temu Kembali Informasi dengan Metode *Vector Space Model*. Jurnal Sistem Informasi Bisnis 02. Semarang: Universitas Stikubank.

[14] Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: *Cambridge University Press*, 2008.

[15] C. H. Yutika, A. Adiwijaya, and S. Al Faraby, "Analisis Sentimen Berbasis Aspek pada *Review Female Daily* Menggunakan TF-IDF dan *Naïve Bayes*," J. Media Inform. Budidarma, vol. 5, no. 2, p. 422, 2021, doi: 10.30865/mib.v5i2.2845.

[16] Muriyatmoko, D., Taufiqurrahman, T., & Humam, A. (2022). Analisis Sentimen Masyarakat Terhadap Konflik Rusia dan Ukraina Menggunakan Metode Naïve Bayes pada Media Sosial Twitter. *Metik Jurnal*, *6*(2), 140–145. https://doi.org/10.47002/metik.v6i2.375