ANALISIS SENTIMEN TERHADAP PROGRAM NATURALISASI TIMNAS INDONESIA PADA *X* MENGGUNAKAN ALGORITMA *NAIVE BAYES*

1st Ariyo Sheva Adhityas
Prodi Sistem Informasi
Universitas Telkom Purwokerto
Purwokerto, Indonesia
21103003@ittelkom-pwt.ac.id

2nd Khairun Nisa Meiah Ngafidin, S.Pd., M.Kom.
Prodi Sistem Informasi
Universitas Telkom Purwokerto
Purwokerto, Indonesia
nisa@ittelkom-pwt.ac.id

***Abstrak* —** Perkembangan era digital saat ini menjadikan media sosial platform utama untuk berinteraksi dan berkomunikasi. Salah satu media sosial yang populer adalah *X*, yang sering digunakan oleh masyarakat Indonesia untuk menyampaikan tanggapan dan opini. Pembicaraan tentang program naturalisasi Timnas Indonesia telah menjadi bahasan menarik di media sosial *X*. Topik terkait program naturalisasi Timnas Indonesia ramai diperbincangkan lagi sejak bulan Februari 2024 setelah perhelatan Piala Asia 2023 sampai Juni 2024 pada saat Timnas Indonesia sedang melakoni kualifikasi Piala Dunia Zona Asia. Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap opini masyarakat mengenai program naturaliasi Timnas Indonesia di media sosial *X*. Data dikumpulkan dengan teknik *crawling* dengan rentang waktu dari bulan Februari hingga Juni 2024 dan diproses melalui tahapan *text preprocessing*, Penelitian ini menggunakan algoritma *Naive Bayes* untuk klasifikasi sentimen menjadi positif, netral, dan negatif. Melakukan balancing data untuk mengatasi ketidakseimbangan data dengan metode *SMOTE*. Evaluasi dilakukan menggunakan *Confusion Matrix* untuk memperoleh nilai *accuracy, precision, recall,* dan *f1-score*. Penelitian ini berhasil mencapai akurasi sebesar 71.11%. Hasil ini menunjukkan bahwa algoritma *Naive Bayes* efektif untuk menganalisis sentimen. Diharapkan dengan adanya penelitian ini dapat mendukung pengambilan keputusan strategis pada program naturalisasi oleh pihak terkait seperti pihak federasi sepakbola Indonesia, untuk meningkatkan kualitas sepakbola Indonesia.

***Kata kunci*— Analisis Sentimen, *Naive Bayes*, Naturalisasi, Timnas Indonesia, *X*.**

I. PENDAHULUAN

Kemajuan teknologi informasi digital telah menjadikan media sosial sebagai platform utama dalam memfasilitasi komunikasi yang cepat dan mudah. Jika dahulu interaksi dilakukan secara langsung, melalui telepon umum, atau surat menyurat, kini media sosial memungkinkan masyarakat berbagi informasi, menyampaikan pendapat, serta membentuk opini publik secara luas. Platform seperti *Facebook, X, Instagram,* dan *TikTok* menjadi sumber utama bagi banyak orang dalam memperoleh informasi secara cepat dan akurat. Peran media sosial semakin signifikan, terutama saat suatu peristiwa menjadi perhatian publik. Dengan pesatnya perkembangan teknologi dan akses internet yang semakin mudah, penggunaan media sosial terus meningkat secara global[1]. Media sosial merupakan salah satu teknologi informasi paling populer untuk berinteraksi, dengan berbagai manfaat seperti memperkuat komunikasi, mendorong kolaborasi, dan membangun komunitas. Salah satu platform yang paling banyak digunakan untuk kebebasan berpendapat adalah *X* [2].

X merupakan platform media sosial yang banyak digunakan oleh berbagai kalangan, termasuk instansi pemerintah, sektor swasta, dan masyarakat umum. Melalui fitur *tweet*, pengguna dapat menyampaikan opini, berbagi pandangan, serta memperoleh informasi dari pengguna lain. Menurut data *We Are Social,* pada Oktober 2023, jumlah pengguna *X* di Indonesia mencapai sekitar 27,5 juta [3]. Dengan tingginya jumlah pengguna *X* di Indonesia, respons masyarakat terhadap berbagai topik di platform ini sangat aktif, terutama terkait program naturalisasi Timnas Indonesia. Perbincangan mengenai isu ini meningkat sejak Februari, pasca Piala Asia 2023, hingga Juni saat Kualifikasi Piala Dunia Zona Asia Putaran Kedua berlangsung. Naturalisasi pemain menjadi sorotan publik karena memicu beragam pendapat, baik yang mendukung maupun mengkritik. Perdebatan ini berkaitan dengan identitas, kebijakan olahraga, serta dampaknya terhadap perkembangan pemain lokal. Di media sosial, opini publik terbagi antara yang menganggap naturalisasi sebagai strategi meningkatkan kualitas tim dan yang khawatir akan berkurangnya peluang bagi pemain lokal untuk berkembang [4].

Program naturalisasi menjadi bagian dari upaya pemerintah dalam meningkatkan sepak bola nasional, sejalan dengan Instruksi Presiden Nomor 3 Tahun 2019 tentang Percepatan Pembangunan Sepakbola Nasional. PSSI selaku federasi sepak bola Indonesia tidak membedakan pemain naturalisasi, yang memiliki hak setara sebagai WNI dan wajib mematuhi peraturan. Naturalisasi diharapkan dapat meningkatkan kualitas serta performa Timnas Indonesia di level internasional, terbukti dengan naiknya peringkat FIFA dari 172 ke 134 dunia. Meski memberi dampak positif, pembinaan pemain muda dan pengembangan liga yang berkualitas tetap perlu diperhatikan [5]. Menanggapi tren yang berkembang saat ini, analisis sentimen dikenal juga dengan sebuah analisis opini atau pengkajian opini yang sangat penting dalam area studi *Natural Language*

Processing (NLP) yang dirancang untuk menganalisis sentimen dan pandangan dari teks secara otomatis. Analisis sentimen merupakan hal yang penting untuk perkembangan kecerdasan buatan [6].

Tanggapan dan opini dari *tweet* pengguna *X* dapat diklasifikasikan melalui analisis sentimen, yang membandingkan opini positif, netral, atau negatif. Proses ini melibatkan pengumpulan, analisis, serta ekstraksi data tekstual terkait suatu fenomena atau tren tertentu [7]. Algoritma *Naive Bayes* dapat digunakan dalam analisis sentimen untuk menentukan apakah suatu opini bersifat positif, netral, atau negatif. Keunggulannya terletak pada kemudahan dan efisiensinya, terutama dalam menangani dataset berdimensi besar dengan performa yang cepat. Dengan mengasumsikan independensi antar fitur, model ini dapat dilatih dengan cepat meskipun data pelatihan terbatas. Meskipun asumsi tersebut tidak selalu sesuai dengan kondisi nyata, *Naive Bayes* tetap memberikan hasil yang baik dan sering dijadikan sebagai pendekatan awal dalam klasifikasi [8].

Tanggapan dan opini masyarakat mengenai program naturalisasi Timnas Indonesia dari media sosial *X* yang berisikan kumpulan *tweet* tersebut diklasifikasi memuat tiga kategori sentimen yaitu negatif, netral, ataupun positif dan kemudian dilakukan penerapan algoritma *Naive Bayes* untuk mengetahui nilai akurasinya [9]. Berdasarkan penjelasan latar belakang tersebut, penelitian ini akan melakukan analisis sentimen data *tweet* pengguna dari *X* dengan topik program naturalisasi Timnas Indonesia.

II. KAJIAN TEORI

A. Analisis Sentimen

Analisis sentimen yaitu proses perupayaan identifikasi dan kategori emosi dari sentimen yang diekspresikan dalam teks. Adanya pertumbuhan yang pesat dari media sosial, semakin banyak opini dari para penggunanya dan hal ini menjadikan analisis sentimen sebagai teknik yang krusial untuk memahami sentimen publik dari berbagai topik dan tren tertentu. Data teks yang telah didapatkan tersebut nantinya akan menjalani beberapa rangkaian seperti preprocessing, ekstraksi, dan klasifikasi. Dalam hal ini teks tersebut akan dapat di kategorikan menjadi positif, negatif, atau netral [10].

B. Naturalisasi

Naturalisasi merupakan proses hukum yang memungkinkan seseorang yang merupakan warga negara asing untuk mendapatkan kewarganegaraan Indonesia. Negara berhak menetapkan peraturan agar seluruh upaya perpindahan kewarganegaraan tidak mengganggu stabilitas negara [11].

C. *X*

X atau nama baru dari *Twitter* ialah *platform* dari media sosial yang memungkinkan para pengguna untuk dapat saling bertukar pesan atau informasi yang sering disebut juga dengan *tweet*. Tepat pada tanggal 22 Juli 2023 *Twitter* berganti nama menjadi *X* setelah berganti kepemilikan. Selain itu, logo tersebut juga berganti menjadi logo *X* putih dengan latar hitam [12].

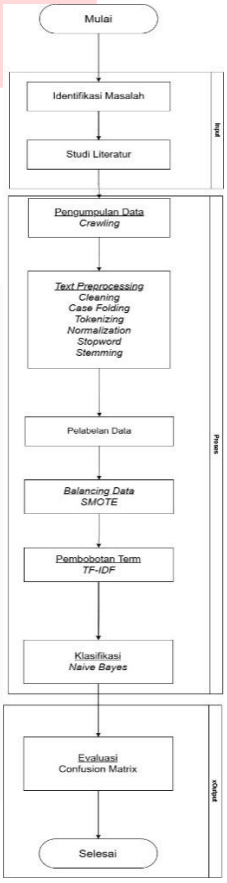
D. *Naive Bayes*

Naive Bayes ialah sebuah algoritma untuk klasifikasi probabilitas dan statistik yang mengasumsikan bahwa setiap atribut bersifat independen, atau karakteristik suatu kelas tidak terkait dengan kelas lainnya. Penggunaan *Naive Bayes* mendasar dengan banyaknya dataset untuk digunakan, sehingga memerlukan sebuah algoritma dengan kinerja klasifikasi yang cepat dan akurasi yang relatif tinggi [13].

E. Google Colab

Google Colab merupakan *tools* yang berbentuk *cloud* yang disediakan oleh *Google*. Pertumbuhan bahasa pemrograman *Python* inilah yang membuat *Google* menghasilkan *Integrated Environment Development* (IDE) atau istilah yang lebih diketahui dengan nama Google Colab. Jenis lingkungan yang dikenakan ialah Jupyter menggunakan ekstensi *file* **.ipynb*. Selaku informasi *python* memiliki berbagai lingkungan untuk program itu sendiri dari *IDLE* yang telah lama dikenakan hingga Spyder dengan lingkungan yang lebih lengkap. Karakter yang basis nya *web*, *Jupyter Notebook* lebih disukai *Google* [14].

III. METODE



Gambar 1. Diagram Alir Penelitian

A. Identifikasi Masalah

Identifikasi masalah menjadi tahap awal yang akan menjadi dasar dalam penelitian ini. Topik yang dikutip dari penelitian yaitu program naturalisasi Timnas Indonesia dan ini akan menjadi garis dasar penelitian ini. Penelitian ini nantinya akan melakukan analisis sentimen terhadap respon mengenai program naturalisasi Timnas Indonesia berdasarkan tweet pengguna X menggunakan algoritma Naive Bayes. Identifikasi masalah digunakan untuk menjadi dasar bagi penelitian yang dijalankan.

B. Studi Literatur

Studi Literatur menjadi tahap selanjutnya dalam penelitian ini. Peneliti mencari referensi berdasarkan jurnal

yang relevan dengan penelitian ini. Pada tahap ini juga akan didapatkan melakukan perbandingan antara penelitian terdahulu dengan penelitian yang akan dijalankan.

C. Pengumpulan Data

Pengumpulan data dapat dijalankan dengan tahap crawling di media sosial *X* menggunakan *Tweet Harvest*. Data yang diambil berupa teks cuitan pengguna dengan kata kunci 'program naturalisasi Timnas Indonesia'. Data yang diambil berbahasa Indonesia dan diambil dalam rentang waktu bulan Maret 2024 hingga Juni 2024.

D. *Text Preprocessing*

Tahap ini memiliki tujuan untuk melakukan pembersihan suatu kata yang kurang dibutuhkan dan kata yang tidak bermakna sama sekali untuk setiap *tweet* data. *Preprocessing* sangat penting dalam analisis teks karena membantu meningkatkan kualitas data, sehingga lebih mudah dipahami dan diolah oleh model analitik atau algoritma *machine learning*. Tahap text preprocessing dilakukan dengan tahapan seperti:

1. *Cleaning*

   Tahapan yang dilaksanakan dengan tujuan untuk menghapus tanda baca, nomor, *link*, simbol, tagar, *username* atau *mention*, *retweet*, dan karakter yang kurang dibutuhkan.

2. *Case Folding*

   Tahapan dengan tujuan untuk menggantikan semua huruf kapital teks menjadi kecil secara keseluruhan.

3. *Tokenizing*

   Tahapan berikutnya yakni *tokenizing* dengan tujuan untuk menguraikan kalimat dalam sebuah teks menjadi sebuah kata per kata.

4. *Normalization*

   Tahapan ini bertujuan untuk mengubah kata-kata yang tidak baku menjadi baku serta memperbaiki kesalahan ejaan atau singkatan, sehingga menghasilkan kata-kata yang lebih tepat dan sesuai kaidah.

5. *Stopword*

   Tahapan yang bertujuan untuk penghapusan sebuah kata yang tidak mempunyai arti penting.

6. *Stemming*

   Tahapan ini dilakukan untuk membenahi sebuah kata kembali bentuk kata asalnya dengan menghapus kata imbuhan di depan dan belakang dari setiap kata.

E. Pelabelan Data

Proses pelabelan data dikategorikan memuat tiga kelas yaitu positif, negatif, dan netral. Data yang dilabelkan berfungsi sebagai dataset untuk proses klasifikasi dengan *Naive Bayes*. Proses pelabelan ini dilakukan dengan menggunakan pendekatan *lexicon-based*, di mana setiap kata dalam teks dianalisis berdasarkan kamus kata atau lexicon yang telah dilengkapi dengan nilai sentimen tertentu. Setiap kata akan diberi label sesuai dengan kategori sentimennya, apakah positif, negatif, atau netral, berdasarkan bobot atau nilai yang ada dalam kamus tersebut.

F. *Balancing Data SMOTE*

Balancing Data menggunakan metode *SMOTE* untuk mengatasi ketidakseimbangan pada data. Metode ini melakukan sintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset melalui interpolasi, dengan cara menciptakan contoh baru yang merupakan kombinasi dari sampel minoritas yang ada. Proses ini membantu meningkatkan representasi kelas minoritas dalam dataset, yang pada gilirannya dapat meningkatkan performa model dalam memprediksi kelas yang kurang terwakili.

G. Pembobotan *Term TF-IDF*

Pembobotan *Term* dengan menggunakan metode *TF-IDF* digunakan untuk menghitung *Term Frequency* dan *Inverse Document Frequency*. Penghitungannya dengan cara membandingkan antara frekuensi sebuah *term* dengan nilai maksimal dari frekuensi *term* pada dokumen tersebut. *IDF* atau *Inverse Document Frequency* adalah perhitungan bagaimana *term* didistribusikan pada dokumen, yang bertujuan untuk memberi bobot lebih tinggi pada kata-kata yang jarang muncul di seluruh koleksi dokumen, namun sering muncul dalam dokumen tertentu. Rumus dari *TF-IDF* yakni sebagai berikut:

$$W_{dt} = TF_{dt} \times IDF_{ft} + 1 \quad (1)$$

W_{dt} : bobot dokumen ke-d terhadap kata ke=t
TF_{dt} : banyaknya kata yang dicari pada sebuah dokumen
IDF_{ft} : Inversed Document Frequency $\left(log\left(\frac{N}{df}\right)\right)$
N : total dokumen
df : banyak dokumen yang memuat kata yang dicari

G. Klasifikasi *Naive Bayes*

Tahapan klasifikasi dengan menerapkan algoritma *Naive Bayes* mengakar pada *Teorema Bayes* dengan menggunakan metode statistik dan probabilitas. Penelitian ini menggunakan salah satu varian *Naive Bayes* yang populer, yakni *Multinomial Naive Bayes*, yang dirancang khusus untuk mengolah data seperti frekuensi kata dalam dokumen. *Multinomial Naive Bayes* berfokus pada penghitungan probabilitas kondisi berdasarkan distribusi frekuensi kata dalam dokumen, yang sangat cocok untuk teks yang berbasis kata-kata atau fitur kategorikal lainnya. Persamaan Teorema Naive Bayes:

$$P(x|z) = \frac{P(x)P(z|x)}{P(z)} \quad (2)$$

IV. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Pada penelitian ini, data yang dikumpulkan melalui *crawling data*, yang merupakan proses pengumpulan informasi secara otomatis dari berbagai sumber di internet, khususnya dari media sosial seperti *X*. Dengan menggunakan *Tweet Harvest,* data yang terkumpul sebanyak 1519 baris data, yang berisi informasi relevan yang akan digunakan untuk analisis lebih lanjut dalam penelitian ini. Metode *crawling* dan penggunaan *Tweet Harvest* ini memungkinkan penelitian untuk mengakses data dalam jumlah besar dan memperoleh informasi yang relevan sesuai dengan tujuan

penelitian, yaitu analisis sentimen atau klasifikasi teks dari media sosial. Hasil dari crawling data terdapat pada Gambar 2 berikut ini.

	created_at	username	full_text
0	Wed Jun 19 21:00:37 +0000 2024	MSNindonesia	Pakar Asal Inggris Ungkap Kesulitan Vietnam un...
1	Wed Jun 19 19:24:06 +0000 2024	okezonenews	3 Pemain Naturalisasi yang Pindah Klub Setelah...
2	Wed Jun 19 18:01:41 +0000 2024	PlatW08	Calon naturalisasi timnas Indonesia
3	Wed Jun 19 16:24:44 +0000 2024	ajimuhammad354	Tren Bloke Core: Cara Mudah Tampil Keren denga...
4	Wed Jun 19 16:23:55 +0000 2024	ajimuhammad354	Kick Off Malam di Indonesia: Lebih Nyaman Bagi...
...	...	...	...
1514	Tue Mar 26 03:02:14 +0000 2024	tvOneNews	Julukan Unik Para Naturalisasi Baru Timnas Ind...
1515	Tue Mar 26 01:28:53 +0000 2024	BolaSportcom	Ragnar Oratmangoen menjadi salah satu nama bar...
1516	Tue Mar 26 01:28:45 +0000 2024	tribunSUPERBALL	Ragnar Oratmangoen menjadi salah satu nama bar...
1517	Mon Mar 25 21:04:47 +0000 2024	nafritzicosta	@SerieA_ID Timnas belanda pun pake naturalisas...
1518	Mon Mar 25 16:36:48 +0000 2024	mudours	Indonesia akhirnya mengecap buah manis dari ki...

Gambar 2. Crawling Data

Pada Gambar 2 tersebut terdapat dataset yang telah dikumpulkan yang berisi tanggal *tweet* dibuat (*created_at*), *username*, dan teks *tweet* (full_text). Perlu diketahui bahwa meskipun rentang waktu yang ditentukan untuk crawling data adalah dari 1 Februari hingga 20 Juni 2024, data yang terkumpul dimulai pada bulan Maret 2024. Hal ini disebabkan oleh faktor limitasi karena pengumpulan seperti ini sudah dibatasi oleh pihak *X* yang menyebabkan proses pengumpulan *tweet* baru didapatkan pada bulan Maret. Meskipun demikian, data yang terkumpul tetap mencakup periode yang relevan dengan topik naturalisasi timnas Indonesia, termasuk berbagai peristiwa penting yang terjadi pada rentang waktu tersebut.

B. *Preprocessing*

Tahap *preprocessing data* dilakukan dengan beberapa proses, yakni *cleaning*, *case folding*, *tokenization*, *stopword*, dan *stemming*. Hasil data setelah *preprocessing* berisi 1519 data yang sudah dibersihkan dari elemen yang tidak diperlukan dan siap digunakan untuk analisis dengan algoritma *Naive bayes*.

a. *Cleaning*

Tahapan yang dilaksanakan dengan tujuan untuk menghapus tanda baca, nomor, link, simbol, tagar, username atau mention, retweet, dan karakter yang kurang dibutuhkan. Hasil dari *cleaning* terdapat pada Gambar 3 berikut.

cleaning
Pakar Asal Inggris Ungkap Kesulitan Vietnam un...
Pemain Naturalisasi yang Pindah Klub Setelah R...
Calon naturalisasi timnas Indonesia
Tren Bloke Core Cara Mudah Tampil Keren denga...
Kick Off Malam di Indonesia Lebih Nyaman Bagi...
...
Julukan Unik Para Naturalisasi Baru Timnas Ind...
Ragnar Oratmangoen menjadi salah satu nama bar...
Ragnar Oratmangoen menjadi salah satu nama bar...
Timnas belanda pun pake naturalisasi dari indo...
Indonesia akhirnya mengecap buah manis dari ki...

Gambar 3. Hasil Cleaning

b. *Case Folding*

Tahapan dengan tujuan untuk menggantikan semua huruf kapital teks menjadi kecil secara keseluruhan. Proses ini dilakukan untuk meningkatkan konsistensi data dan menghilangkan perbedaan antara huruf besar dan huruf kecil yang dapat mempengaruhi analisis. Hasil dari *case folding* terdapat pada Gambar 4 berikut.

case_folding
pakar asal inggris ungkap kesulitan vietnam un...
pemain naturalisasi yang pindah klub setelah r...
calon naturalisasi timnas indonesia
tren bloke core cara mudah tampil keren denga...
kick off malam di indonesia lebih nyaman bagi...
...
julukan unik para naturalisasi baru timnas ind...
ragnar oratmangoen menjadi salah satu nama bar...
ragnar oratmangoen menjadi salah satu nama bar...
timnas belanda pun pake naturalisasi dari indo...
indonesia akhirnya mengecap buah manis dari ki...

Gambar 4. Case Folding

c. *Tokenizing*

Tahapan berikutnya yakni tokenizing dengan tujuan untuk menguraikan kalimat dalam sebuah teks menjadi sebuah kata per kata. Hasil dari *tokenizing* terdapat pada Gambar 5 berikut.

tokenize
[pakar, asal, inggris, ungkap, kesulitan, viet...
[pemain, naturalisasi, yang, pindah, klub, set...
[calon, naturalisasi, timnas, indonesia]
[tren, bloke, core, cara, mudah, tampil, keren...
[kick, off, malam, di, indonesia, lebih, nyama...
...
[julukan, unik, para, naturalisasi, baru, timn...
[ragnar, oratmangoen, menjadi, salah, satu, na...
[ragnar, oratmangoen, menjadi, salah, satu, na...
[timnas, belanda, pun, pake, naturalisasi, dar...
[indonesia, akhirnya, mengecap, buah, manis, d...

Gambar 5. Tokenizing

d. *Normalization*

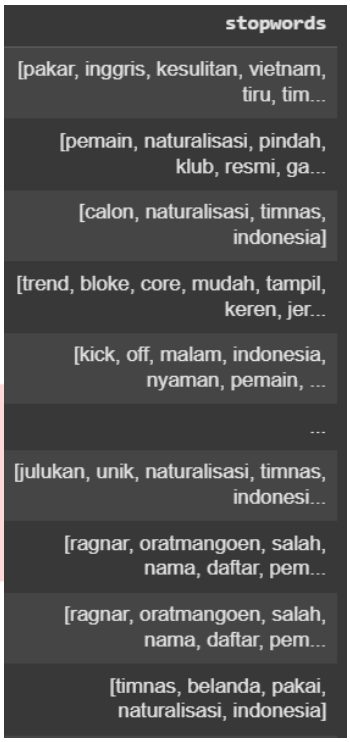
Tahapan yang bertujuan untuk mengganti kata-kata tidak baku menjadi bentuk baku dan memperbaiki kesalahan ejaan pada kata-kata yang disingkat atau tidak sesuai dengan kaidah bahasa. Hasil dari *normalization* terdapat pada Gambar 6 berikut.

normalization
[pakar, asal, inggris, ungkap, kesulitan, viet...
[pemain, naturalisasi, yang, pindah, klub, set...
[calon, naturalisasi, timnas, indonesia]
[trend, bloke, core, cara, mudah, tampil, kere...
[kick, off, malam, di, indonesia, lebih, nyama...
...
[julukan, unik, para, naturalisasi, baru, timn...
[ragnar, oratmangoen, menjadi, salah, satu, na...
[ragnar, oratmangoen, menjadi, salah, satu, na...
[timnas, belanda, pun, pakai, naturalisasi, da...

Gambar 6. Normalization

e. *Stopword*

Tahapan yang bertujuan untuk penghapusan sebuah kata yang tidak mempunyai arti penting seperti kata penghubung, kata depan, dan kata sambung. Hasil dari *stopword* terdapat pada Gambar 7 berikut.



stopwords
[pakar, inggris, kesulitan, vietnam, tiru, tim...
[pemain, naturalisasi, pindah, klub, resmi, ga...
[calon, naturalisasi, timnas, indonesia]
[trend, bloke, core, mudah, tampil, keren, jer...
[kick, off, malam, indonesia, nyaman, pemain, ...
...
[julukan, unik, naturalisasi, timnas, indonesi...
[ragnar, oratmangoen, salah, nama, daftar, pem...
[ragnar, oratmangoen, salah, nama, daftar, pem...
[timnas, belanda, pakai, naturalisasi, indonesia]

Gambar 7. Stopword

f. *Stemming*

Tahapan ini dilakukan untuk membenahi sebuah kata kembali bentuk kata asalnya dengan menghapus kata imbuhan di depan dan belakang dari setiap kata. Hasil dari *stemming* terdapat pada Gambar 8 berikut.

	stemming_data
0	British Expert Difficult Vietnam Imitate Indon...
1	Playing Naturalization Moving the Official Clu...
2	Indonesian national team naturalization candidate
3	Trend Bloke Core Easily Looks Cool Favorite So...
4	Kick off at night Indonesia is comfortable pla...
...	...
1507	Unique Nickish Naturalization of Indonesian Na...
1508	Ragnar Oratmangoen One of the names of the Ind...
1509	Ragnar Oratmangoen One of the names of the Ind...
1510	The Dutch national team uses Indonesian natura...
1511	Indonesia Soy Sauce Sweet Fruit Gait Hunter Er...

Gambar 8. Stemming

C. Pelabelan Data

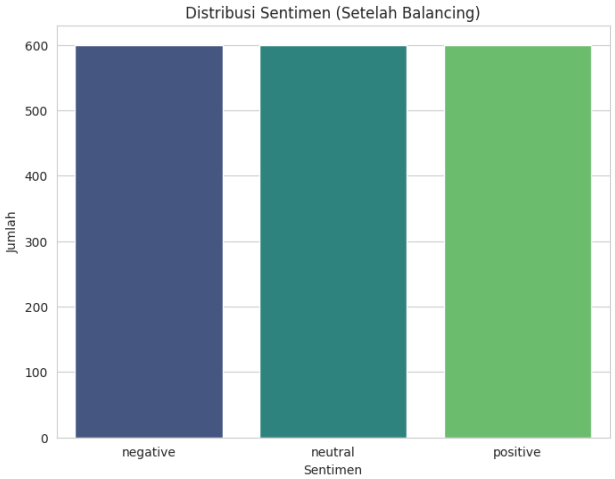
Pelabelan data ini dilakukan dengan menggunakan teks yang telah menjalani proses preprocessing. Pelabelan dilakukan menggunakan pendekatan lexicon based yakni InSet Lexicon (Indonesia Sentiment Lexicon) dikenal sebagai metode pendekatan kamus dengan pemberian nilai pada setiap kata dalam suatu opini menurut kamus bobot

penilaian positif dan negatif. InSet Lexicon adalah sebuah kamus sentimen yang secara khusus dikembangkan untuk analisis sentimen dalam bahasa Indonesia, di mana setiap kata dalam kamus ini diberi bobot berdasarkan polaritas sentimennya, apakah itu positif, negatif, atau netral.

	stemming	sentiment
0	pakar inggris sulit vietnam tiru timnas indone...	negative
1	main naturalisasi pindah klub resmi gabung tim...	neutral
2	calon naturalisasi timnas indonesia	neutral
3	trend bloke core mudah tampil keren jersey bol...	neutral
4	kick off malam indonesia nyaman main naturalis...	negative

Gambar 9. Hasil Pelabelan Data

Meskipun proses pelabelan data telah dilakukan untuk menentukan polaritas sentimen, langkah selanjutnya yang sangat penting adalah *balancing data*, karena sering kali menghadapi ketidakseimbangan antara jumlah data untuk setiap kelas sentimen (positif, negatif, netral). Ketidakseimbangan ini dapat menyebabkan model klasifikasi memiliki bias terhadap kelas yang lebih dominan, sehingga hasil prediksi menjadi kurang akurat, terutama dalam memprediksi kelas yang minoritas. Salah satu teknik yang efektif dalam mengatasi ketidakseimbangan data adalah *SMOTE* (Synthetic Minority Over-sampling Technique). *SMOTE* bekerja dengan cara membuat contoh sintetik dari kelas minoritas dengan melakukan interpolasi antar data yang ada.



Gambar 10. Diagram Label Setelah SMOTE

Pada Gambar 10 tersebut distribusi data sentimen setelah dilakukan proses balancing menggunakan metode *SMOTE*. Dalam gambar tersebut, terlihat bahwa jumlah data pada setiap kelas, yaitu negative, neutral, dan positive, telah dibuat seimbang, masing-masing memiliki sekitar 600 data. Penyeimbangan ini memastikan bahwa tidak ada kelas yang mendominasi dalam dataset, sehingga model klasifikasi *Naive Bayes* dapat mempelajari pola dari setiap kelas secara adil dan merata.

Dengan dataset yang telah diseimbangkan, kemungkinan bias terhadap kelas mayoritas dapat diminimalkan, sehingga model dapat bekerja dengan lebih optimal dalam memprediksi sentimen pada data baru.

D. Pembobotan *Term TF-IDF*

Proses selanjutnya yaitu pembobotan term menggunakan TF-IDF. Metode ini akan memberikan nilai penting pada setiap kata dalam dataset berdasarkan seberapa sering kata tersebut muncul di setiap dokumen (TF) dan frekuensi kemunculan kata dalam keseluruhan dataset (IDF). Terdapat contoh dokumen TF-IDF pada Tabel 1 berikut ini.

Tabel 1. Dokumen TF-IDF

Dokumen	**Teks**
1	tolong naturalisasi sty biar timnas indonesia maju
2	coba naturalisasi wasit indonesia keren
3	timnas naturalisasi klok doang yak turun Indonesia

Selanjutnya, menghitung frekuensi kata yang sering muncul dalam setiap dokumen yang disebut *Term Frequency* (TF). Hasil dari Term Frequency dapat dilihat pada Tabel 2 berikut ini.

Tabel 2. Term Frequency

Term	TF		
	D1	D2	D3
tolong	1	0	0
naturalisasi	1	1	1
sty	1	0	0
biar	1	0	0
timnas	1	0	1
indonesia	1	1	1
maju	1	0	0
coba	0	1	0
wasit	0	1	0
keren	0	1	0
klok	0	0	1
doang	0	0	1
yak	0	0	1
turun	0	0	1

Setelah menghitung TF, selanjutnya menghitung *Inverse Document Frequency* (IDF), yaitu frekuensi kemunculan kata dalam seluruh dokumen. IDF dihitung sebagai logaritma dari rasio antara total jumlah dokumen dalam koleksi dengan jumlah dokumen yang mengandung kata tersebut. Hasil IDF dapat dilihat pada Tabel 3 berikut ini.

Tabel 3. Inverse Document Frequency

Term	**DF**	**D/df**	**IDF**
tolong	1	7	0.845
naturalisasi	3	2.3	0.368
sty	1	7.00	0.845
biar	1	7	0.845
timnas	2	3.5	0.544
indonesia	3	2.3	0.368
maju	1	7	0.845

Term	DF	D/df	IDF
coba	1	7	0.845
wasit	1	7	0.845
keren	1	7	0.845
klok	1	7	0.845
doang	1	7	0.845
yak	1	7	0.845
turun	1	7	0.845

Kemudian, setiap nilai TF dikalikan dengan nilai IDF untuk mendapatkan nilai TF-IDF. Nilai TF-IDF dapat dilihat pada Tabel 4 berikut.

Tabel 4. TF-IDF

Term	TF-IDF		
	D1	D2	D3
tolong	0.845	0	0
naturalisasi	0.368	0.368	0.37
sty	0.845	0.000	0.000
biar	0.845	0	0
timnas	0.544	0	0.54
indonesia	0.368	0.37	0.37
maju	0.845	0	0
coba	0.000	0.85	0
wasit	0.000	0.85	0.000
keren	0.000	0.85	0
klok	0.000	0	0.85
doang	0.000	0	0.845
yak	0.000	0	0.85
turun	0	0.000	0.85

E. *Data Split*

Tahap selanjutnya adalah tahap pembagian data, di mana dataset akan dibagi menjadi 2 bagian yaitu *data training* dan *data testing*. Pembagian akan dilakukan dengan perbandingan 90% *data train* dan 10% *data test.* Hasil pembagian data didapatkan 1620 baris *data train* dan 180 baris *data test*. Menggunakan perbandingan 90:10 karena untuk menjaga keseimbangan antara pelatihan model dan evaluasi performa. Dengan 90% data train, model mendapatkan lebih banyak data untuk belajar dan mengenali pola atau hubungan antar fitur dalam dataset. Pembagian *Data Split* dapat dilihat pada Gambar 11 berikut.

```
x_train =  1620
x_test =  180
y_train =  1620
y_test =  180
```

Gambar 11. Pembagian Data Split

F. Klasifikasi *Naive Bayes*

Proses selanjutnya yaitu klasifikasi dengan *Naive Bayes*. Pada Klasifikasi Naive Bayes ini menggunakan MultinomialNB yang ditunjukkan pada Gambar 12 berikut.

```
MultinomialNB
MultinomialNB()
```

Gambar 12. Klasifikasi MultinomialNB

MultinomialNB merupakan varian dari algoritma *Naive Bayes* dan dianggap paling cocok dalam hal klasifikasi pada teks. Pada *Multinomial Naive Bayes*, setiap kata dalam teks dianggap sebagai fitur independen yang tidak memiliki keterkaitan satu sama lain, sehingga algoritma ini mampu menghitung probabilitas berdasarkan frekuensi kemunculan kata dalam dokumen.

G. Evaluasi Model

Evaluasi Model dilakukan dengan menggunakan Confusion Matrix dengan menghitung accuracy, precision, recall, dan f1-score. Hasil dari evaluasi model dapat dilihat pada Gambar 13 berikut.

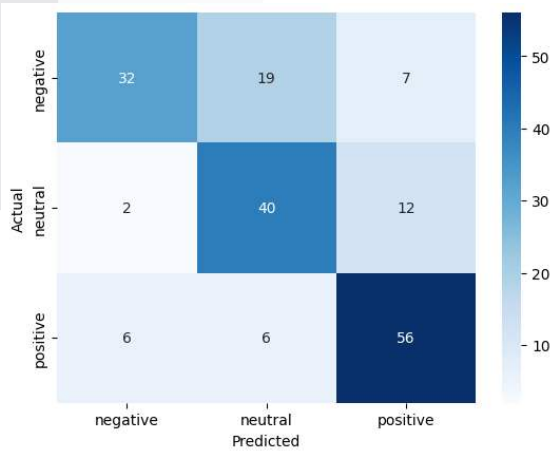
```
Accuracy: 0.7111111111111111
Precision: 0.7244672364672365
Recall: 0.7111111111111111
F1 Score: 0.707993407153071
              precision    recall  f1-score

    negative       0.80      0.55      0.65
     neutral       0.62      0.74      0.67
    positive       0.75      0.82      0.78

    accuracy                           0.71
   macro avg       0.72      0.71      0.70
weighted avg       0.72      0.71      0.71
```

Gambar 13. Hasil Akurasi

Setelah melakukan proses menghitung nilai dalam kinerja keseluruhan sistem, didapatkan tingkat akurasi sebesar 71%. Output dari gambar 13 juga memberikan hasil dari precision, recall, dan f1-score dengan masing-masing sebesar 72%, 71%, dan 70%. Pada proses berikutnya akan menampilkan bentuk visualisasi dan hasil angka *confusion matrix*.



Gambar 14. Heatmap Confusion Matrix

Gambar 14 tersebut menunjukkan heatmap hasil dari performa model klasifikasi dengan tiga kelas yaitu; negative, neutral, dan positive. Heatmap merupakan representasi visual dari Confusion matrix yang digunakan untuk memahami

performa model klasifikasi dengan lebih mudah. Warna pada heatmap menunjukkan intensitas jumlah data, dengan warna lebih gelap menunjukkan jumlah yang lebih besar.

Pada garis vertikal terdapat data actual, sementara pada garis horizontal merupakan prediksi pada model. Setiap kotak menunjukkan jumlah data dalam kategori. Model berhasil memprediksi 32 data negative secara benar, namun salah memprediksi 2 data neutral dan 6 data positive sebagai negative. Untuk kelas neutral, model memprediksi 40 data dengan benar, tetapi salah memprediksi 19 data sebagai negative dan 12 data sebagai positive. Pada kelas positive, model memprediksi 56 data dengan benar, namun salah memprediksi 6 data sebagai neutral dan 7 data sebagai negative. Warna pada heatmap menunjukkan intensitas jumlah data, dengan warna lebih gelap menunjukkan jumlah yang lebih besar.

H. Hasil Analisis

Pada penelitian ini, melakukan pengujian untuk menganalisis hasil akurasi *Naive Bayes* terkait *tweet* topik Naturalisasi Timnas Indonesia. Penelitian ini membandingkan antara tiga sentimen yakni; positif, negatif, dan netral. Setelah menganalisis semua *tweet*, mendapatkan hasil bahwa Sebagian besar memiliki sentimen negatif. Dari *tweet* yang dianalisis, terdapat 1512 *tweet* yang telah dihapus duplikat dan sudah melewati tahap dari *preprocessing*. Setelah melakukan proses pelabelan, diperoleh hasil bahwa terdapat 605 tweet yang memiliki label sentimen negatif, sementara 499 tweet dikategorikan sebagai sentimen positif, dan sebanyak 408 tweet memiliki label sentimen netral. Klasifikasi ini memungkinkan analisis lebih mendalam terkait persebaran sentimen dalam data, sehingga dapat memberikan gambaran yang lebih jelas mengenai kecenderungan opini yang terdapat dalam tweet yang telah diproses. Setelah melewati proses Balancing data menggunakan SMOTE diperoleh hasil accuracy sebesar 71%, precision 72%, recall 71%, dan f1-score 70%.

Berdasarkan hasil analisis tersebut, sebelum dilakukan balancing data terdapat gambaran umum mengenai program naturalisasi ini belum dapat diterima dengan baik oleh masyarakat dengan lebih banyak opini yang berlabel negatif daripada positif atau netral. Tetapi, ada masyarakat yang telah menerima program naturalisasi ini dengan contoh terdapat opini seperti "Ikut menyemangati dan menghargai para pemain Timnas Sepak bola indonesia yg berasal dari pemain naturalisasi yg saat ini sdg ikhlas berjuang membela Indonesia Terimakasih utk kalian semua kalian semua adalah saudara kami bangsa indonesia". Hasil dari sentimen yang diperoleh dapat digunakan oleh pihak federasi sepakbola Indonesia sebagai evaluasi untuk memahami opini dari masyarakat secara lebih mendalam. Dengan demikian, federasi sepakbola Indonesia dapat mengoptimalkan strategi dalam mengelola program naturalisasi, memperkuat dukungan dari masyarakat, dan memaksimalkan kontribusi program tersebut terhadap kemajuan Timnas Indonesia di level internasional.

V. KESIMPULAN

Penelitian ini menunjukkan algoritma Naive bayes efektif untuk menganalisis sentimen terkait program naturalisasi Timnas Indonesia, dengan akurasi sebesar 71%. Implementasi algoritma Naive Bayes terkait program naturalisasi Timnas Indonesia di media sosial X menunjukkan hasil yang cukup baik, didukung oleh tahapan preprocessing dan melakukan penyeimbangan data menggunakan SMOTE guna mengatasi ketidakseimbangan kelas, serta evaluasi model dengan Confusion Matrix yang membuktikan bahwa algoritma ini mampu mengklasifikasikan sentimen dengan mendapatkan hasil yang cukup baik. Hasil dari klasifikasi analisis sentimen pada penelitian ini menunjukkan bahwa lebih banyak opini negatif yang diberikan pada topik Naturalisasi Timnas Indonesia. Hasil akhir menunjukkan bahwa 605 opini data berlabel negatif dan sebanyak 499 opini berlabel positif kemudian untuk opini netral terdapat sebanyak 408 data.

REFERENSI

[1] T. Maura Safa Ramadhanti, R. Br Tarigan, A. Fatahilla, D. Ramadhan Rangkuti, and M. Fharisi, "Media Sosial dan Pembentukan Opini Publik," Jurnal Komunikasi, Sosial, dan Ilmu Politik, no. 3032–7482, pp. 67–74, Jan. 2025.

[2] F. Zaini, J. W. Sari, and F. N. Hasan, "Analysis of Public Sentiment Related to The Failure of Indonesia to Host U-20 Using Multinomial Naïve Bayes Classifier," Jurnal Teknik Informatika (Jutif), vol. 4, no. 6, pp. 1409–1418, Dec. 2023.

[3] N. A. Azmi, A. T. Fathani, D. P. Sadayi, I. Fitriani, and M. R. Adiyaksa, "Social Media Network Analysis (SNA): Identifikasi Komunikasi dan Penyebaran Informasi Melalui Media Sosial Twitter," Jurnal Media Informatika Budidarma, vol. 5, no. 4, p. 1422, Oct. 2021.

[4] M. Ghafur Rahman Lubis, D. Sambora Sitompul, T. Muhammad Giovanni, F. Ramadhani, and S. Dewi, "Evaluasi Kinerja Algoritma Support Vector Machine (SVM) Dalam Analisis Sentimen Publik Terhadap Naturalisasi Timnas Indonesia di Twitter," *JALAKOTEK: Journal of Accounting Law Communication and Technology*.

[5] Randi Ilham. (Apr. 29, 2024). Seberapa Berdampak Naturalisasi Pemain Terhadap Kemajuan Sepak Bola Indonesia [Kumparan]. Available: https://kumparan.com/randi-ilham/seberapa-berdampak-naturalisasi-pemain- terhadap-kemajuan-sepak-bola-indonesia-21tfqqBMCtw/2

[6] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, p. 102048, Apr. 2024.

[7] F. P. P. Subandi, F. Romadlon, I. Nurisusilawati, and A. Chindyana, "Sentiment Analysis of Indonesian Interest in Korean Food Based on Naïve Bayes Algorithm," Jurnal Sosioteknologi, vol. 21, no. 3, pp. 337–346, Dec. 2022.

[8] M. Artur, "Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross validated selection of the best number of features," in Procedia Computer Science, Elsevier B.V., pp. 564–570, Jul. 2021.

[9] Afandi R, Hanif F, and Hasan N, "Analisis Sentimen Opini Masyarakat Terkait Penyelenggaraan Sistem Elektronik Menggunakan Metode Logistic Regression," 2022.

[10] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," Applied Sciences (Switzerland) vol.13, no.7, Apr. 01, 2023.

[11] A. Hakim Zuryat, "Perlindungan Hak Kewarganegaraan Berdasarkan Asas Persamaan Derajat Dalam Hal Naturalisasi Para Pemain Sepakbola Indonesia Berdasarkan Undang-Undang No.12 Tahun 2006 Tentang Kewarganegaraan," 2020.

[12] S. Wahyu, "Perbandingan Model Algoritma Klasifikasi Pada Analisis Sentimen Opini Masyarakat Terhadap Layanan Kereta Cepat Jakarta Bandung (The Whoosh)". Konferensi Nasional Ilmu Komputer (KONIK), 2023.

[13] A. Mutia Mantika, A. Triayudi, and R. T. Aldisa, "Sentiment Analysis on Twitter Using Naïve Bayes and Logistic Regression for the 2024 Presidential Election," 2024.

[14] R. Gelar Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," Jurnal Teknologi Dan Sistem Informasi Bisnis, vol. 5, no. 1, pp. 55–60, Feb. 2023.