

PENERAPAN ALGORITMA *NAÏVE BAYES* PADA TWITTER UNTUK ANALISIS SENTIMEN MASYARAKAT TERHADAP KETUA UMUM PSSI TERPILIH PERIODE 2023-2027 BAPAK ERICK THOHIR

Feroli Bahru Al Rouf^{1*}, Grace Gata², Gandung Triyono³, Nawindah⁴

^{1, 2} Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia

Email: ¹ferolibahru19@gmail.com, ²grace.gata@budiluhur.ac.id, ³gandung.triyono@budiluhur.ac.id, ⁴nawindah@gmail.com
(* : corresponding author)

Abstrak-Persatuan Sepak Bola Seluruh Indonesia (PSSI), sebagai organisasi yang menaungi sepak bola di Indonesia, menggelar Kongres Luar Biasa (KLB), Kamis, 16 Februari 2023 di Hotel Shangrila, Jakarta. Kongres Luar Biasa yang di laksanakan ini sehubungan dengan tragedi Kanjuruhan. Ketua Umum sebelumnya resmi mengakhiri masa baktinya. Masa jabatan kepengurusan lama seharusnya berakhir pada November 2023, namun di percepat akibat tragedi berdarah tersebut. Agenda Kongres Luar Biasa ini akan memilih kepengurusan PSSI periode 2023 – 2027. Berdasarkan hasil yang di gelar tersebut, terpilih Ketua Umum PSSI baru periode 2023 – 2027. Hal tersebut sering dibicarakan di kalangan masyarakat, salah satunya di media sosial Twitter. Pengguna Twitter akan memberikan komentar dan opini positif atau negatif mengenai ketua umum PSSI baru. Oleh karena itu diperlukannya analisis sentimen masyarakat guna menyelaraskan dan memberi pandangan baru mengenai suatu isu yang berkaitan tentang Erick Thohir dengan menggunakan metode algoritma *Naïve Bayes* untuk mengetahui kecenderungan sentimen masyarakat tentang isu tersebut di twitter. Penelitian ini menerapkan metodologi *Cross-Industry Standard Process for Data Mining (CRISP-DM)*. Subjek yang diambil menggunakan *API twitter* dengan *tools rapidminer* sebanyak 3667 tweet menggunakan kata kunci “ErickThohirPSSI” pada 10 Juni sampai 18 Juni 2023, proses dimulai dengan *crawling*, menentukan sentimen positif atau negatif, *preprocessing*, hingga didapatkan hasil akhir. Setelah melewati beberapa proses maka didapatkan 1368 tweet yang terdiri dari 1143 dengan sentimen positif, 42 dengan sentimen negatif dan 183 dengan sentimen netral. Penelitian ini menghasilkan tingkat *accuracy* sebesar 57.22%, *precision* 88.06% dan *recall* 59.74% dengan perbandingan 60:40 untuk data *training* dan *testing*.

Kata Kunci: *Text Mining, Naïve Bayes, Analisis Sentimen, PSSI, Erick Thohir, CRISP-DM*

APPLICATION OF THE *NAÏVE BAYES* ALGORITHM ON TWITTER FOR ANALYSIS OF PUBLIC SENTIMENT FOR THE ELECTED PSSI CHAIRMAN FOR THE 2023-2027 PERIOD MR. ERICK THOHIR

Abstract-*The Indonesian Football Association (PSSI), as the organization that oversees football in Indonesia, held an Extraordinary Congress (KLB), Thursday, 16 February 2023 at the Shangrila Hotel, Jakarta. This Extraordinary Congress was held in connection with the Kanjuruhan tragedy. The previous Chairperson officially ended his tenure. The term of the old management was supposed to end in November 2023, but was accelerated due to the bloody tragedy. The agenda for this Extraordinary Congress will elect PSSI management for the 2023-2027 period. Based on the results that were held, a new PSSI General Chair was elected for the 2023-2027 period. This is often discussed among the public, one of them on social media Twitter. Twitter users will provide positive or negative comments and opinions regarding the new PSSI chairman. Therefore, it is necessary to analyze public sentiment in order to harmonize and provide new views on an issue related to Erick Thohir by using the *Naïve Bayes* algorithm to determine the tendency of public sentiment about this issue on Twitter. This research applies the *Cross-Industry Standard Process for Data Mining (CRISP-DM)* methodology. Subjects were taken using the Twitter *API* with *rapidminer* tools as many as 3667 tweets using the keyword “ErickThohirPSSI” from 10 June to 18 June 2023, the process began with *crawling*, determining positive or negative sentiment, *preprocessing*, until the final results were obtained. After going through several processes, 1368 tweets were obtained consisting of 1143 with positive sentiment, 42 with negative sentiment and 183 with neutral sentiment. This study produces an accuracy rate of 57.22%, precision 88.06% and recall 59.74% with a ratio of 60:40 for training and testing data.*

Keywords: *Text Mining, Naïve Bayes, Sentiment Analysis, PSSI, Erick Thohir, CRISP-DM*

1. PENDAHULUAN

Kerusuhan antar supporter sepak bola terjadi pada 1 Oktober 2022 di pertandingan Liga 1 Indonesia antara Arema FC vs Persebaya di Stadion Kanjuruhan, Malang. Menyebabkan 712 korban, 132 di antaranya meninggal dunia, 96 orang luka berat dan 484 orang luka ringan atau sedang. Kerusuhan yang terjadi pasca pertandingan di sebabkan karena pihak penyelenggara tidak profesional, tidak saling memahami tugas dan peran serta mengalihkan tanggung jawab kepada pihak lain. Sikap dan praktik seperti itu sudah ada bertahun-tahun dalam penyelenggaraan kompetisi sepak bola nasional. Karena itu, berbagai rekomendasi disampaikan kepada berbagai pihak untuk memperbaiki masa depan sepak bola Indonesia, salah satunya adalah membenahi struktur kepengurusan PSSI[1].

Persatuan Sepak Bola Seluruh Indonesia (PSSI), sebagai organisasi yang menaungi sepak bola di Indonesia, menggelar Kongres Luar Biasa (KLB), Kamis, 16 Februari 2023 di Hotel Shangrila, Jakarta. Kongres Luar Biasa yang di laksanakan ini sehubungan dengan tragedi Kanjuruhan. Ketua PSSI sebelumnya resmi mengakhiri masa bakti sebagai ketua umum PSSI periode 2019 – 2023. Masa jabatan kepengurusan lama seharusnya berakhir pada November 2023, Namun di percepat akibat tragedi berdarah tersebut. Agenda Kongres Luar Biasa ini akan memilih kepengurusan PSSI periode 2023 – 2027. Berdasarkan hasil yang di gelar tersebut, Erick Thohir terpilih sebagai Ketua Umum PSSI periode 2023 - 2027. Sedangkan Ratu Tisha dan Zainudin Amali akan menjadi Wakil Ketua Umum.

Pengguna media sosial semakin meningkat di masyarakat, memengaruhi perkembangan opini publik. Twitter, salah satu jejaring sosial paling populer di dunia, memungkinkan penggunaanya untuk berbagi pesan yang disebut tweet di berbagai platform [2]. Pengguna *Twitter* akan memberikan komentar dan opini terhadap ketua umum PSSI baru. Namun, akan sulit bagi pengguna untuk melihat tweet secara langsung tanpa menandainya sebagai positif, negatif atau netral.

Oleh karena itu, pengguna memerlukan klasifikasi untuk mengetahui apakah tweet memiliki nilai netral, negatif, atau positif. Teknik text mining, analisis sentimen menggunakan algoritma Naive Bayes, dapat digunakan untuk melakukan analisis ini. *Naïve Bayes* adalah algoritma klasifikasi yang tergolong sangat sederhana dan efisien. Ini merupakan metode populer untuk klasifikasi teks dan telah menunjukkan kinerja yang baik. Namun, satu kelemahan yang dimiliki adalah sensitivitas terhadap pemilihan fitur. Jika terlalu banyak fitur digunakan, hal ini tidak hanya meningkatkan waktu perhitungan, tetapi juga dapat mengurangi akurasi klasifikasi[3]. Selain itu, belum adanya penelitian yang mengklasifikasikan sentimen opini masyarakat di *Twitter* terhadap ketua umum PSSI terpilih periode 2023 - 2027 dengan menggunakan algoritma *Naïve Bayes*. Rumusan permasalahan dari penelitian ini adalah bagaimana penerapan algoritma dan nilai akurasi yang diperoleh algoritma *Naïve Bayes* dalam melakukan analisis sentimen masyarakat Indonesia terhadap ketua umum PSSI terpilih. Tujuan penelitian ini adalah untuk untuk memperoleh hasil dari analisis sentimen dengan metode *Naïve Bayes* serta dapat dijadikan masukan bagi PSSI khususnya ketua umum PSSI terpilih periode 2023-2027 yang mendapat respon positif dari masyarakat bahwa masyarakat Indonesia menaruh harapan besar. Dibawah kepemimpinannya yang memiliki pengalaman di industri sepak bola dapat mampu memperbaiki, membuat terobosan dan menransformasi sepak bola Indonesia sehingga bisa maju, berkembang dan bersaing dengan negara lain.

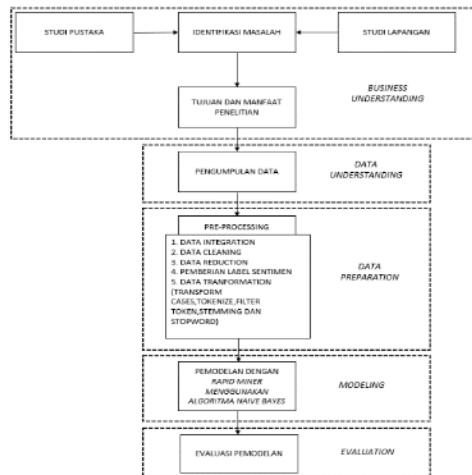
Data *tweet* yang telah dikumpulkan selanjutnya dilakukan analisis lebih lanjut untuk mengetahui secara umum bagaimana tanggapan masyarakat. Salah satu cara untuk memperoleh data-data yang dibutuhkan untuk melakukan analisis sentimen yaitu dengan melakukan *crawling* menggunakan *tools rapidminer*. Tweet dapat di analisis sentimennya dengan cara manual untuk menentukan sentimen positif ataupun negatif, akan tetapi apabila jumlah tweet yang digunakan berjumlah ribuan maka akan memerlukan waktu yang sangat lama. Oleh karena itu digunakan metode CRISP-DM yang akan membuat proses analisis menjadi lebih terstruktur dan sistematis.

Penelitian terdahulu telah dilakukan analisis sentimen terhadap Kinerja PSSI menggunakan algoritma *Support Vector Machine*, *dan Decision Tree*[4]. Telah dilakukan juga penelitian mengenai sentimen terhadap performa timnas sepak bola indonesia dengan algoritma *Naïve Bayes* [5]. Serta penelitian tentang Klasifikasi dan Analisis Sentimen pada Data Twitter Menggunakan Algoritma Naïve Bayes (Studi Kasus: Timnas Indonesia Senior, U-23, dan U-19) menggunakan algoritma *Naïve Bayes* [6]. Maka pada penelitian ini akan menggunakan metode text mining dan analisis sentimen menggunakan algoritma *Naïve Bayes* untuk mengetahui opini masyarakat terhadap ketua umum PSSI periode 2023-2027 sehingga dapat digunakan sebagai data analisis opini publik. Kemudian dari hasil pengukuran yang didapat akan dilakukan pengujian menggunakan *confusion matrix*.

Berdasarkan penjelasan diatas, peneliti akan melakukan analisis sentimen tentang terpilihnya Erick Thohir sebagai ketua umum PSSI baru melalui media sosial *twitter* menggunakan metode algoritma *Naïve Bayes Classifier*. Hasil dari penelitian ini untuk melihat sentimen masyarakat Indonesia di media sosial *Twitter* terkait terpilih-nya Erick Thohir sebagai Ketua Umum PSSI[7].

2. METODE PENELITIAN

2.1 Metode CRISP-DM



Gambar 1. Tahapan Penelitian Menggunakan CRISP-DM

Penelitian ini terdiri dari beberapa langkah seperti pada Gambar 1, Pada penelitian ini, metode *Cross Industry Standard Procedures for Data Mining* (CRISP-DM) akan digunakan. CRISP-DM menggunakan model proses pengembangan data untuk strategi pemecahan masalah.[8]. Berikut ini penjelasan langkah-langkah dalam proses penelitian:

a. *Business Understanding*

Dari permasalahan yang muncul terkait ketua umum PSSI terpilih Erick Thohir peneliti mencoba untuk memahami permasalahan tersebut dengan melakukan mempelajari artikel terkait PSSI dan berita dari internet mengenai ketua umum PSSI Erick Thohir, kemudian mencari tweet atau trending topik di twitter terkait dengan ketua umum PSSI Erick Thohir.

b. *Data Understanding*

Pada tahap penelitian ini, tujuannya adalah untuk memahami data yang akan digunakan dan mengidentifikasi tweet yang relevan dengan topik analisis menggunakan beberapa kata kunci, seperti "Ketua Umum PSSI Erick Thohir". Proses pengumpulan data dilakukan menggunakan *software rapidminer*. Sebelum memulai proses pengambilan data *tweet*, peneliti membuat koneksi antara *RapidMiner* dengan *Twitter*. Selanjutnya, peneliti menentukan periode waktu untuk pengambilan data tweet, yang berlangsung dari tanggal 10 Juni 2023 hingga 18 Juni 2023. Jumlah total data tweet yang terkumpul sebanyak 3.667 tweet[5].

c. *Data Preprocessing*

Pada tahap ini, data yang telah diberi label sentimen, akan dilakukan data *preprocessing*. Berikut adalah proses yang dilakukan selama data *preprocessing*[9]:

1. *Remove Duplicate*

Remove Duplicate adalah proses untuk menghapus sebuah data yang muncul secara berulang.

2. *Replace*

Replace adalah proses untuk menghilangkan kata yang terdapat pada tweets seperti RT(ReTweet), https:, dan @username.

3. *Case Folding*

Case Folding adalah proses mengubah huruf dalam dokumen menjadi bentuk huruf kecil

4. *Tokenizing*

Tokenizing atau tokenisasi adalah suatu proses untuk membagi dokumen menjadi kata-kata yang menyusun dan menghilangkan tanda baca, simbol apapun yang bukan huruf.

5. *Stopword*

Stopword adalah proses menghilangkan sebuah kata yang tidak penting, sehingga kata yang digunakan hanya kata-kata yang penting saja.

6. *Stemming*

Stemming adalah proses mencari akar kata atau pengembalian bentuk kata menjadi bentuk kata dasar dengan menghilangkan imbuhan.

7. *Term Weighting (TF-IDF)*

TF-IDF adalah prosedur pembobotan kata yang melibatkan langkah-langkah sebagai berikut:

- a. Menemukan nilai *Term Frequency* (TF), dengan menghitung jumlah kata yang muncul dalam dokumen
- b. Menemukan nilai *Document Frequency* (DF), dengan menghitung jumlah dokumen yang mengandung sebuah kata
- c. Menemukan nilai *Invers Document Frequency* (IDF), dengan membagi jumlah dokumen yang mengandung sebuah kata (DF)
- d. Menghitung Bobot = *Term Frequency* (TF) x *Invers Document Frequency* (IDF)[4].

d. *Modeling*

Pada tahap ini, peneliti akan memodelkan dataset yang telah melalui proses *preprocessing* sebelumnya. Tahap pemodelan ini melibatkan pembuatan model prediktif untuk memprediksi sentimen dalam dataset. Selanjutnya, dataset akan dibagi menjadi dua bagian, yaitu data latih dan data uji. Penelitian ini akan membagi data dalam dua tahap. Pada tahap pertama, data akan dibagi dengan perbandingan/*split data* antara data latih dan data uji dengan rasio 60:40, 70:30, dan 80:20. Pada tahap kedua, penggunaan *cross-validation* dilakukan dengan membagi data menjadi 10 bagian (*number of folds* = 10). Pada tahap ini, penelitian menggunakan algoritma *Naïve Bayes* untuk menghasilkan nilai sentimen negatif, netral, dan positif[10].

e. *Evaluation*

Setelah tahapan pemodelan selesai, peneliti melakukan pengujian fungsionalitas sistem dengan metode *confusion matrix* mempersentasikan hasil dari evaluasi model dengan menggunakan tabel *matrix*, jika dataset terdapat 3 class, class pertama yang dianggap sebagai *class positive*, class kedua yang dianggap sebagai *class negative* dan class ketiga yang dianggap sebagai *class neutral*. Hasil yang didapatkan menggunakan *confusion matrix* adalah nilai *accuracy*, *precision* dan *recall*. Mengevaluasi metode klasifikasi terhadap hasil dari pemodelan menggunakan *confusion matrix* terhadap algoritma *Naïve Bayes* untuk mengukur performa [11].

2.2 *Naïve Bayes Classifier*

Naïve Bayes adalah algoritma klasifikasi yang sangat sederhana dan efisien. Algoritma *Naïve Bayes* juga merupakan metode yang populer untuk klasifikasi teks dan memiliki kinerja yang baik, namun juga memiliki kelemahan yaitu sangat sensitif terhadap seleksi fitur. Terlalu banyak fitur tidak hanya meningkatkan waktu perhitungan, tetapi juga dapat mengurangi akurasi klasifikasi. Klasifikasi Bayesian, berdasarkan teorema Bayes, mampu mengklasifikasikan dengan akurasi dan kecepatan tinggi ketika diterapkan pada database yang berisi data dalam jumlah besar. Pada algoritma *Naïve Bayes* setiap dokumen di presentasikan dengan memasukkan “a1, a2, a3, ..., an“ dimana a1 adalah kata pertama an berikutnya sampai, sedangkan V yaitu label kategori, selanjutnya yaitu mencari nilai tertinggi dari kategori teks yang diujikan (VMAP) persamaan VMAP yaitu pada rumus sebagai berikut :

$$VMAP = vi \in argmax P_{Vji} \cap P_{AiV} \quad (1)$$

3. HASIL DAN PEMBAHASAN

3.1 Pemahaman Terhadap Bisnis (*Business Understanding*)

Pada tahap pemahaman bisnis ini, peneliti berusaha memahami objek atau masalah tersebut dengan mempelajari artikel yang berkaitan dengan PSSI serta berita online tentang Erick Thohir, ketua umum PSSI. Permasalahan yang diangkat peneliti adalah akurasi algoritma *Naïve Bayes* dalam mengklasifikasikan sentimen tweet yang berkaitan dengan Erick Thohir. Untuk penelitian ini, data dikumpulkan melalui media sosial Twitter. Twitter memiliki keunggulan bahwa jika banyak tweet tentang topik yang sama, topik tersebut otomatis menjadi trending topik, yang memungkinkan peneliti mendapatkan data tweet tentang topik yang mereka butuhkan untuk proses *crawling* data.

3.2 Pemahaman Terhadap Data (*Data Understanding*)

Peneliti mencoba memahami data yang digunakan dalam proses *Data Understanding* ini dengan menemukan *keywords* yang relevan untuk proses *crawling* data pada tweet. Selain itu, penulis melakukan pengecekan topik trending Erick Thohir, yang menjadi topik trending pertama pada 10 Juni 2023 hingga 18 Juni 2023, dengan

menggunakan topik pendukung seperti #ErickThohirPSSI. Data yang dikumpulkan dengan menggunakan *keywords* #ErickThohirPSSI berjumlah 3667 tweet.

3.3 Persiapan Data (*Data Preparation*)

Data yang telah dikumpulkan dan diberi label sentimen kemudian dilakukan *data preprocessing* oleh peneliti pada tahap *data preparation*, yang dibagi menjadi dua tahapan berikut:

- Tahapan *preprocessing*, tahapan ini dilakukan oleh peneliti ketika data *tweet* diberi label sentimen positif, negatif dan netral.
- Penentuan data latih dan data uji: Sebelum masuk ke tahapan *modeling*, proses ini menentukan perbandingan antara data latih dan data uji.

3.3.1 Data Preprocessing/Pra Pemrosesan Data

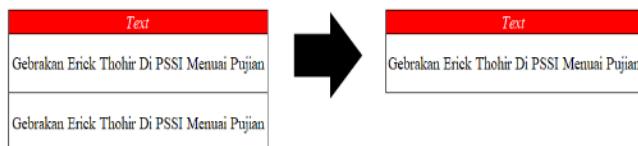
Proses ini meliputi beberapa tahapan, antara lain *data integration*, *replace*, *remove duplicates*, pelabelan sentimen, *case folding*, *tokenize*, *filter tokens*, *stemming*, dan *filter stopwords*. Berikut ini merupakan tahapan-tahapannya.

- Replace**, bertujuan untuk menghapus link, mention, hashtag, dan karakter khusus yang terdapat pada *tweet*. Tahapan ini dilakukan dengan menggunakan operator *replace*.

Tabel 1. Contoh Proses *Replace*

Sebelum Proses <i>Replace</i>	Setelah Proses <i>Replace</i>
RT @Arkaan_ : Erick Thohir mendapatkan apresiasi karena banyak melakukan terobosan selama menjabat ketua umum PSSI #GueBarengErickThohir	Erick Thohir mendapatkan apresiasi karena banyak melakukan terobosan selama menjabat ketua umum PSSI

- Remove duplicates**, yaitu proses menghilangkan data yang terduplikasi. Proses ini dilakukan dengan menggunakan operator *Remove Duplicates*. Dari proses ini jumlah *tweet* berkurang dari 3.667 menjadi 1.368 *tweet*.



Gambar 2. Hasil Proses *Remove Duplicates*

- Pelabelan Sentimen**, proses ini dilakukan dilakukan oleh pakar dengan menganalisa setiap *tweet* yang sudah didapatkan. Label sentimen yang diberikan pada *tweet* berupa sentimen positif, netral, dan negatif.

Tabel 2. Contoh *Tweet* Setelah Diberikan Label Sentimen

Text	Label
Erick Thohir mendapatkan apresiasi karena banyak melakukan terobosan selama menjabat ketua umum PSSI	Positif
Selama ini PSSI diisi oleh mafia pengaturan score wajar selalu terjadi intrik saat ini pun Erick Thohir sedang diincar kesalahannya bila ada pasti diakan terus di cercanya bonus bonus dri Negara sebaiknya diinfestasikan dan masa depan dan asuransi	Netral
Erick thohir blm jadi cawapres udah ngecewakan pecinta bola yg terlanjur beli tiket pertandingan ngurus pssi aja campur bohong gmn urus bumn tuh	Negatif

Kini sebanyak total 1.368 komentar yang menjadi dataset yang berisi 1.143 komentar bernilai positif , 42 komentar bernilai negatif dan 183 komentar bernilai netral. Ini berarti dalam dataset terdapat sebanyak 84% komentar bernilai positif , 3% komentar bernilai negatif dan 13% komentar bernilai netral.

- Transform Case**, merupakan suatu proses yang dilakukan untuk menyeragamkan semua bentuk huruf. Pada penelitian ini bentuk huruf yang digunakan adalah huruf kecil. Sehingga semua bentuk huruf kapital akan diubah menjadi huruf kecil.