Social Media Analysis of the Selection of John Herdman as the Head Coach of the Indonesian National Football Team

Luki Ardiantoro [a], Muhammad Irfan [b] , Kurniawan Dwi Firnanda [a]

[a]*Department of Informatic Engineering, Universitas Bina Sehat PPNI, Mojokerto, Indonesia*
[b]*Department of Informatic Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia*

ARTICLE INFO

Article history:
Received May 3, 2026
Received in revised form
June 10, 2026
Accepted June 25, 2026
Available online June 30, 2026

Keywords:
Football
John Herdman
Naïve Bayes
Rapid Miner
Social media

ABSTRACT

As a popular sport, the position of the Indonesian national football team coach attracts attention. Criticism and comments on their performance are often subjective and lead to controversy. This study conducted an objective performance evaluation of the national team coach based on data from social media. This method is commonly applied to measure public attitudes by classifying opinions into positive or negative categories through text processing of data taken from platforms such as X (Twitter) and YouTube. Based on experiments and tests from a large amount of tweet data from X users, it was found that the highest accuracy value was obtained from the 600th data point, which consisted of 280 positive and 320 negative data points. Based on experiments and tests using the parameters, it was found that the highest accuracy value was in the test with parameter 5, which had an accuracy of 94.170% and an AUC value of 0.686. The important findings are as follows: Public sentiment toward John Herdman is predominantly positive, which reflects satisfaction with leadership expectations and increased trust in team direction.

1 Introduction

The selection of foreign coaches for the Indonesian National Football Team often triggers significant reactions in the public sphere, especially on social media platforms that serve as the main outlet for Indonesian football supporters' expressions[1]. These changing interaction dynamics allow for the formation of massive and rapid public opinion, where virtual movements on social media have proven capable of exerting real pressure on football-related policies in the country. This aligns with the findings that virtual public spaces can generate public opinions that pressure both the government and related organizations[1]. Literature Review Research related to social media analysis in the context of sports has extensively used the sentiment analysis approach to understand public opinion on supporter profiles and market composition[2].

The background of the problem is that the use of technology in football has rapidly developed in many countries. The quantitative approach provides a more accurate method for assessing a team's strengths and weaknesses. As a popular sport, the position of the national team coach attracts attention. Criticism and comments on their performance are often subjective and lead to controversy. Therefore, the researchers believe that it is necessary to conduct an objective performance evaluation of the national team coach based on data. The trend in football research emphasizes data-driven performance analysis, tactical context, and game complexity [3]. Big data opens up opportunities for tactical analysis in football but demands the integration of sports science and contextual interpretation[2].

The problem formulation is how a quantitative approach based on data mining can be used to objectively analyze the selection of the national team coach to foster supporters'support and affection toward the national team. With high supporter support, a positive atmosphere is created, and the market value of the national team increases. High supporter support will automatically create a positive atmosphere, and the market value of the national team will increase. This is economically beneficial for the national team's commercial activities, such as advertising, apparel, product promotion, television broadcasting rights, branding, and market value.

The urgency of this research is twofold: 1) The utilization of data-based technology has become an important factor in improving sports performance in various countries. However, its application in Indonesia, particularly for understanding formations and the dynamics of player roles, is still not optimal and is partial in nature. 2) The assessment of the national team's coach's performance is generally based on final results and subjective opinions, which often leads to controversy and does not provide a basis for an objective and sustainable evaluation.

This research process refers to the six stages outlined in CRISP-DM, which are explained as follows: Research data refer to information that is collected and used in the context of scientific research. Research data can consist of facts, numbers, statistics, statements, observations, and other information relevant to the topic being studied. These data can be obtained through various methods, such as surveys, experiments, interviews, observations, literature reviews, or secondary data collection. Research data play an important role in the scientific research process. The data collected must be accurate, reliable, and relevant to the research questions. In addition, research data must be organized, analyzed, and interpreted to generate information and knowledge that can answer research questions, test hypotheses, and reveal patterns, trends, or relationships in the phenomena being investigated. In this study, public data were used.

2 Literature Study

Football requires teamwork and team cohesion; therefore, player position formations are used to facilitate assignments and role-sharing during matches. The coach's role as a strategist is crucial in determining the direction of the game and the team's harmony. The Indonesian national team appointed John Herdman on January 13, 2026. John Herdman is an English professional football coach who currently serves as the head coach of the Indonesian national team. He is widely known for his contributions to the Canadian national team as the first coach to lead both the women's and men's teams from the same country to the World Cup. The response of supporters regarding the selection of the national football team coach and how supporters feel satisfied with reality. This serves as a fulfillment of expectations, with social media becoming a public opinion barometer. Based on the national team's performance in each

Corresponding Author: ipan.ardianto@gmail.com

Doi : 10.30996/elsains.v8i1.133496

match, goals scored or conceded, and even failures, heated discussions can be sparked [3].

The selection of foreign coaches for the Indonesian National Football Team often triggers significant reactions in the public sphere, especially on social media platforms that serve as the main outlet for Indonesian football supporters' expressions[1]. These changing interaction dynamics allow for the formation of massive and rapid public opinion, where virtual movements on social media have proven capable of exerting real pressure on football-related policies in the country. This aligns with findings that virtual public spaces can generate public opinions that pressure both the government and related organizations[1][3].

The level and enthusiasm of the public to know all of this can be measured using a sentiment analysis. Sentiment analysis is a powerful tool for identifying incoming sentiment patterns. By understanding public sentiment, if it has a positive trend, then supporter support is guaranteed to increase. In the long term, this will create mass loyalty in supporting the national teams. Indirectly, the team's morale will increase, which will affect the national team's performance in achieving success.

The Naïve Bayes Classifier has advantages such as being simple, fast, and highly accurate. The Naïve Bayes Classifier method for text classification uses the attributes of words that appear in a document as the basis for classification. A study shows that although the assumption of independence between words in a document cannot be fully met, the performance of the Naïve Bayes Classifier in classification is relatively very good. In this study, several issues were identified, including the lack of classification of positive and negative sentiments regarding John Herdman and the unknown accuracy of the Naïve Bayes Classifier algorithm in classifying test data for sentiment analysis [4]. For future research, the use of additional algorithms, such as support vector machines or random forests, can refine sentiment classification results, especially with diverse and large datasets. Furthermore, exploring advanced preprocessing techniques tailored for Bahasa Indonesia, alongside expanding the training dataset, could further enhance the model performance and generalizability [4][5]. This would enable a more nuanced understanding of the complex linguistic structures and contextual inferences inherent in social media discourse, thereby improving the robustness of sentiment analysis models[6][7]. Moreover, incorporating dynamic feature extraction methods that adapt to evolving linguistic patterns within social media could offer a significant boost in classification accuracy[8][9].

The integration of deep learning architectures, such as Bidirectional Encoder Representations from Transformers or Long Short-Term Memory networks, might further refine the capture of semantic nuances and long-range dependencies often present in social media texts, thereby surpassing the limitations of traditional machine learning approaches. Finally, addressing the challenge of duplicate data, as noted in research on presidential candidate sentiment analysis, is crucial for improving data effectiveness and model accuracy [10]. Moreover, exploring the semantic relationships between words and incorporating sentiment lexicons specifically tailored for Indonesian slang could mitigate misclassifications, as demonstrated in studies addressing the complexities of social media language [7]. Such advancements in methodology would not only improve the precision of sentiment analysis but also provide a more comprehensive understanding of public discourse surrounding significant events, such as the appointment of a national sports team coach. Future research could also focus on refining sentiment analysis models to account for the dynamic and often colloquial nature of Indonesian social media discourse[11][12]. Furthermore, the integration of multimodal data, including images and videos often associated with social media posts, has the potential to deepen the analysis by providing additional contextual cues that influence public sentiment [13][12].

Additionally, leveraging advanced deep learning methods, such as BiLSTM, and integrating large pretrained language models could further enhance analytical depth, particularly for large datasets[14]. Furthermore, incorporating context-specific adjustments to capture nuanced language, including sarcasm and domain-specific jargon, can refine sentiment classification[2][15]. Future work could also investigate incorporating a neutral sentiment category to provide a more comprehensive understanding of public perception, moving beyond a simple positive or negative dichotomy[16]. Given the complexities of language, especially in informal social media contexts, future research should increasingly explore deep learning techniques, particularly transformer-based models, for their superior ability to manage such intricacies in sentiment analysis[17]. Developing models with manual annotations can improve accuracy by capturing subtle human language nuances[18].

Moreover, a richer lexicon that includes non-standard Indonesian and English terms, alongside the translation of emojis and emoticons, would minimize out-of-vocabulary words and enhance the applicability of models such as multilingual BERT to social media texts[11][5]. Further advancements could involve the development of hybrid models that combine the strengths of traditional machine learning with deep learning approaches, potentially leading to more robust and accurate sentiment detection across diverse Indonesian social media content[19].

Furthermore, the inclusion of multimodal data, such as audio and video from social media posts, could provide a more holistic understanding of user expressions, thereby enriching sentiment analysis[20][21]. In addition, future studies could investigate transfer learning techniques within AI frameworks for sports analysis to better capture intricate sentiment patterns in social media data on football coaches and national team events, as demonstrated in the analyses of Indonesian public reactions to coaches like Shin Tae-yong[22][23]. This could involve leveraging advanced techniques such as BERT for nuanced language understanding [12] and exploring dynamic sentiment analysis to capture evolving public opinion [24][25]. Additionally, implementing network analysis alongside sentiment classification could reveal influential users and communities shaping public discourse around coaching appointments[26].

Moreover, expanding the evaluation to include additional datasets, such as multilingual corpora or domain-specific texts in healthcare, finance, or legal domains, could enhance the model’s applicability and generalizability across diverse use cases[26]. Furthermore, the application of XAI techniques can provide crucial transparency into model predictions, particularly in online safety systems, where decisions impact user moderation[24]. Additionally, incorporating manual labeling into future research endeavors would significantly enhance the accuracy of the analysis, particularly when benchmarked against Indonesian lexicons [12].

Further refinements could involve experimenting with sophisticated pre-trained models, such as mBERT or XLM-R, which leverage multilingual understanding and broader linguistic nuances for improved performance[5]. This approach capitalizes on advancements in natural language processing to better discern

Corresponding Author: ipan.ardianto@gmail.com
Doi: 10.30996/elsains.v8i1.133496

the intricate sentiment shifts and cultural specificities often present in Indonesian social media discussions concerning public figures. Such a comprehensive approach would provide a more granular understanding of public sentiment, moving beyond simplistic positive-negative categorizations to reveal underlying emotional complexities[25][27].

3 Methodology

This research uses the Naïve Bayes algorithm with data from Twitter collected using the hashtags "John Herdman" and "Timnas Indonesia." The research stages refer to the Cross Industry Standard for Data Mining (CRISP-DM) method, which consists of the following six stages: 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) modeling, 5) evaluation, and 6) deployment, as shown in Fig. 1.

At this stage, the approach used was the Cross Industry Standard for Data Mining (CRISP-DM) method. CRISP-DM is a method that employs a data development process model that is widely used by experts to solve problems. This research process refers to the six stages present in CRISP-DM, which are as follows:

1. **In terms of business Understanding**, this study analyzed public sentiment toward John Herdman, the coach of the Indonesian national football team. The purpose of the research is to determine public sentiment toward John Herdman's leadership. Data were collected from social media X using Google Colab and the RapidMiner application. After the data were collected, several processes were carried out to process the data, including replacing RT, links, hashtags, mentions, and symbols, and removing duplicates to produce tweets ready for classification analysis using the Naïve Bayes algorithm.
2. **Data Understanding Based on the collected data**, the next step is to understand the data to achieve the research objectives. The data collected from X using Google Colab were further analyzed to ensure their suitability and relevance to the research objectives.
3. Data preparation encompasses all processing steps to build a dataset that will be applied to the classification modeling tool. This data preparation is intensive, with processing activities to ensure that the data are ready for use. The stages of data preparation included:
   i. Data Collection, collecting data from X using the keywords "John Herdman," "STY," and "Timnas."
   ii. Data Cleansing, cleaning the data from special characters, links, punctuation, and other unnecessary elements, including stopwords such as "and," "or," and "also" that are not relevant to sentiment analysis.
   iii. Tokenization is the process of breaking down the text into smaller units (words or phrases) for further analysis.
   iv. Stemming or Lemmatization, returning words to their base form, either through stemming (trimming word endings) or lemmatization (returning words to their base form according to context).
   v. Removing Noise, deleting numbers, symbols, and excessively repeated words in the text to avoid disrupting the analysis.
   vi. Normalization, performing text normalization, such as removing punctuation or replacing certain words with standard representations.
   vii. Sentiment Labeling Marking data with the appropriate sentiment, such as "positive" or "negative." This labeling can be performed manually or automatically using a sentiment dictionary or classification model. After the data preparation stage was complete, the next step was to perform sentiment analysis using the Naïve Bayes Algorithm Classification.
4. **Modeling**. At this stage, a classification method using the Naïve Bayes algorithm was applied. The data were divided into three main attributes: tweets, positive, and negative. RapidMiner was used to categorize the data according to the desired classification.
5. **Evaluation**: The evaluation stage aims to analyze and measure the accuracy of the modeling that has been performed. The evaluation was conducted to ensure that the model created was appropriate and accurate for this research case. The evaluation results determine the next steps, whether the research can continue, or needs to be repeated if it does not align with the initial plan.
6. **Deployment**, the final stage, is to disseminate the research results, which are then compiled in the form of a report or presentation, to convey the findings and knowledge obtained from the modeling and evaluation of this data mining process.

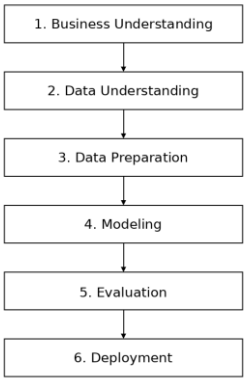

Figure 1. Crisp-DM methodology

Research data were obtained by scraping the social media platform X using keywords relevant to the research topic. A total of 2,000 tweets were collected in the initial stage. Subsequently, duplicate removal, text cleaning, Indonesian language selection, and relevance filtering were performed, resulting in 1,200 tweets that were suitable for analysis. From this dataset, 600 tweets were selected for manual sentiment labeling into positive, negative, and neutral categories. The labeled dataset was then split using the hold-out method, with 80% allocated for the training data (480 samples) and 20% for the testing data (120 samples). This step was performed to ensure the quality of the data used in the development and evaluation of the classification model. Table 1 shows the data research dataset processing workflow.

The data used in this research were qualitative, which characterizes or describes phenomena that can be observed and recorded. Qualitative data are divided into two types:

1. Primary Data are obtained directly from the original source, allowing researchers to observe and record information directly from the research object. The technique for collecting primary data is through direct observation, which enables researchers to obtain comprehensive data and information related to the object.

2. Secondary Data, which are data that already exist or have been processed previously and obtained from other sources as additional information. Techniques for collecting secondary data include literature studies, which involve collecting data from literature, reading, noting, and processing research materials. Textbooks Collecting information through books relevant to the research topic.

Table 1. Research Dataset Processing Workflow

No	Data Processing Stage	Process Description	Amount of Data	Eliminated Data
1	Data Scraping	Data retrieval from social media X with API/scraper based on research keywords	2,000	-
2	Duplication Removal	Deleting tweets that have identical content or the same retweets	1,850	150
3	Data Cleaning	Remove URLs, mentions, hashtags, emojis, special characters, and empty data	1,720	130
4	Language Selection	Select only Indonesian language tweets	1,500	220
5	Relevance Selection	Deleting tweets that are not relevant to the research topic	1,200	300
6	Sentiment Labeling	Manual labeling into positive, negative, and neutral	600	600*
7	Final Data Ready for Analysis	Completely labeled dataset	600	-
8	Training Data (80%)	Used for model training	480	-
9	Testing Data (20%)	Used for model testing	120	-

Figure 2 shows the Python (Google Colab) pipeline for this research, which includes scraping tweets from X (Twitter), preprocessing (cleaning, tokenizing, stemming, normalization), sentiment labeling (simple lexicon-based), and export to CSV (ready for RapidMiner).

```python
query = "John Herdman OR Timnas Indonesia lang:id since:2026-01-16 until:2026-05-20
max_tweets = 5000

tweets = []

for i, tweet in enumerate(sntwitter.TwitterSearchScraper(query).get_items()):
    if i >= max_tweets:
        break
    tweets.append([tweet.date, tweet.content])

df = pd.DataFrame(tweets, columns=['date', 'tweet'])

print("Total tweets collected:", len(df))
df.head()
```

Figure 2. Scrape tweets from X

Figure 3 shows the data cleansing process for identifying and correcting or removing incorrect, corrupted, inaccurate, incomplete, or irrelevant data from a dataset. This step is crucial for improving data quality, ensuring consistency, and preparing raw data for accurate analyses or machine learning modeling.

```python
def clean_tweet(text):
    text = text.lower()
    text = re.sub(r'http\S+', '', text)  # remove links
    text = re.sub(r'@\w+', '', text)     # remove mentions
    text = re.sub(r'#\w+', '', text)     # remove hashtags
    text = re.sub(r'rt[\s]+', '', text)  # remove RT
    text = re.sub(r'[^a-zA-Z\s]', '', text)  # remove symbols/numbers
    text = text.strip()
    return text

df['clean_text'] = df['tweet'].apply(clean_tweet)
```

Figure 3. Data cleansing

The RapidMiner workflow is shown in Figure 4. The RapidMiner workflow (often called a “process” in the software) is a visual, drag-and-drop representation of a data science or machine learning pipeline. It allows users to design end-to-end analytics from data ingestion to model deployment without writing code, using a series of connected blocks known as “operators.”

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="10.3.000">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="10.3.000" expanded="true" name="Process">

    <!-- Load CSV Dataset -->
    <operator activated="true" class="read_csv" compatibility="10.3.000" expanded="true" height="68" name="Read
      <parameter key="file" value="C:\data\tweets_herdman.csv"/>
      <parameter key="column_separators" value=","/>
      <parameter key="first_row_as_names" value="true"/>
      <parameter key="encoding" value="UTF-8"/>
    </operator>

    <!-- Set Role (Label) -->
    <operator activated="true" class="set_role" compatibility="10.3.000" expanded="true" height="68" name="Set
      <parameter key="attribute_name" value="label"/>
      <parameter key="target_role" value="label"/>
```

Figure 4. RapidMiner work flow

Table 2. Sentiment Labeling Guidelines

Label	Criteria	Sample of Tweet
Positive	Contains opinions, support, appreciation, or good assessments of the topics discussed	This coaching program is very helpful for European players to perform well.
Neutral	Informative in nature, does not show any particular emotion or opinion	The major tournament is still yet to take place in January 2026.
Negative	Contains criticism, complaints, dissatisfaction, or negative assessment of the topic	Ah, at most, PSSI and this coaching selection are clearly ineffective. It would be better if STY called him back.

Sentiment labeling was performed manually on 600 tweets that underwent preprocessing and relevance selection. The annotation process was carried out by three annotators knowledgeable about sentiment analysis and natural language processing (two IT lecturers and one IT student). Each tweet was

Corresponding Author: ipan.ardianto@gmail.com
Doi: 10.30996/elsains.v8i1.133496

categorized into one of three sentiment classes: positive, negative, or neutral (Table 2).

A positive label was assigned to tweets expressing support, appreciation, or favorable assessments of the research topic. A negative label was assigned to tweets containing criticisms, complaints, dissatisfaction, or unfavorable assessments. A neutral label was assigned to tweets that merely conveyed information without indicating a specific sentiment. The labeling process was conducted independently by two annotators. In the event of a discrepancy in the annotation results, a discussion was held with a third annotator acting as an arbiter to determine the final label. To measure labeling consistency, the Inter-Annotator Agreement (IAA) was assessed using Cohen's Kappa. A Cohen's kappa value of ≥ 0.80 indicated a high level of agreement, meaning that the labeling results were considered reliable for use in training and testing the sentiment classification model.

Table 3. Splitting the Training and Testing Datasets

No	Sentiment Label	Training (80%)	Testing (20%)	Total
1	Positive	176	44	220
2	Neutral	168	42	210
3	Negative	136	34	170
	Total	480	120	600

Table 2 shows the sentiment labeling guidelines, and the obtained sentiment labels are presented in Table 3. The accuracy of a Naive Bayes classifier is the proportion of correct predictions (both true positives and true negatives) made by the model out of the total number of cases examined. It measures the performance of unseen data. It is commonly used for evaluating classification tasks, such as spam filtering, where high accuracy (often >90%) can be achieved. The accuracy is formulated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where:
TP (True Positives) = correctly predicted positive cases
FN (False Negatives) = actual positives that the model missed

The precision of a Naive Bayes classifier measures the accuracy of positive predictions and is defined as the ratio of true positives to all predicted positives. This indicates how many predicted spam emails (for example) were actually spam. High precision means that the model has a low false-positive rate, which is often critical in text classification and medical diagnosis. The formula is:

$$Precission = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The next metric in the Naïve Bayes measurement is recall. Recall measures ability to detect actual positives. Recall in the context of the Naive Bayes classifier is a performance metric that measures how well the model identifies all the actual positive cases:

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

High recall → The model captures most of the positive cases (e.g., detects almost all spam emails or positive sentiments)
Low recall → The model misses several positive cases.

4 Discussion

Research data refers to information collected and used in the context of scientific research. Research data can include facts, figures, statistics, statements, observations, and other information relevant to the topic being studied. These data can be obtained through various methods, such as surveys, experiments, interviews, observations, literature analysis, or secondary data collection. Research data play an important role in the scientific research process. The data collected must be accurate, reliable, and relevant to the research question. Additionally, the research data must be organized, analyzed, and interpreted to generate information and knowledge that can answer the research question, test hypotheses, and reveal patterns, trends, or relationships in the phenomena being studied. In this study, public data were used.

Based on experiments and tests from a large amount of tweet data from X users, it was found that the highest accuracy value was obtained from the 600th data point, which consisted of 280 positive and 320 negative data points. Based on experiments and tests using parameters, it was found that the highest accuracy value was in the test with parameter 5, which had an accuracy of 94,17% and an AUC value of 0.686 (see Fig. 6). The implementation of Naïve Bayes can conduct sentiment analysis on public opinion regarding the New Normal issue in Indonesia. The testing results on the training and testing data ratio show the best ratio of 70% and 30%, with precision of 92,86%, and recall values of 94.55%. The model also determines the accuracy levels of positive and negative sentiments toward public comments about John Herdman on social media X using a dataset of 600 entries. The findings of this study can be used to observe the dominant sentiment classification of public satisfaction with Coach John Herdman on social media. Figure 5 shows examples of raw data results that are ready for RapidMiner.

```
id,tweet,clean_text,tokenized,stemmed,label

1,"John Herdman bagus banget strategi nya buat Timnas Indonesia","john herdman bagus strategi timnas Indonesia","[john,herdman,bagus,strategi,timnas,indonesia]", "john herdman bagus strategi timnas indonesia","positive"
2,"Gaya main timnas jadi lebih rapi sejak dilatih Herdman","gaya main timnas lebih rapi sejak latih herdman","[gaya,main,timnas,rapi,latih,herdman]","gaya main timnas rapi latih herdman","positive"
3,"Saya kurang yakin dengan John Herdman sebagai pelatih","kurang yakin john herdman pelatih","[kurang,yakin,john,herdman,pelatih]","kurang yakin john herdman latih","negative"
```

Figure 5. Raw data in *.csv format example

The data for this research was obtained from social media/forum X, using the hashtags "John Herdman" and "Timnas Indonesia." After collecting the relevant tweets, Google Colab was used to retrieve the data, followed by processing in RapidMiner with the Google Colab X extension to facilitate the extraction and management of the tweet data. The next stages included data description, attribute selection, and cleansing. The data were then exported to CSV format for easier analysis. The collected data were obtained from tweets between January 16, 2026, and May 20, 2026. Test data = 600 tweets, and the confusion matrix is shown in Table 4.

Table 4. Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	260 (TP)	15 (FN)
Actual Negative	20 (FP)	305 (TN)

$$Accuracy = \frac{260 + 305}{260 + 305 + 20 + 15} = \frac{565}{600}$$

$$Accuracy = 94.17\ \%$$

The confusion matrix provides a more detailed understanding of the classification performance than accuracy alone.

- True Positives (TP = 260): Positive sentiments toward John Herdman that were correctly classified as positive.
- True Negatives (TN = 305): Negative sentiments that were correctly classified as negative.
- False Positives (FP = 20): Negative sentiments that were incorrectly classified as positive.
- False Negatives (FN = 15): Positive sentiments that were incorrectly classified as negative.

In this study, the Naïve Bayes algorithm achieved an accuracy of 94.17% in classifying public sentiment toward Herdman. This level of accuracy indicates that this method is effective in identifying positive or negative public opinions, especially on social media platforms.

Precision measures the correctness of positive predictions and is calculated using the following formula:

$$Precission = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$= 260 / 280$$

$$= 92{,}86\ \%$$

In the context of this Naive Bayes classifier, recall is a performance metric that measures how well the model identifies all actual positive cases. There is:

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$= 260 / 275$$

$$= 94{,}55\ \%$$

The performance of the Naïve Bayes classification model was evaluated using the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various classification thresholds, providing a comprehensive view of the model performance independent of a specific decision threshold.

Additional Evaluation Metrics: F1-Score and Confusion Matrix Analysis. In addition to accuracy, precision, and recall, this study evaluated the model using the F1-score, which is the harmonic mean of precision and recall. The F1-score is particularly important when dealing with datasets that are not perfectly balanced because accuracy alone may provide an overly optimistic view of the model performance.

The F1-score is calculated as follows:

$$F1 = 2\times(\text{Precision}\times\text{Recall}) / (\text{Precision}+\text{Recall})$$

Using the values obtained in this study:

Precision = 92.86% (0.9286)
Recall = 94.55% (0.9455)

$$F1 = 2\times(0.9286\times0.9455)/(0.9286+0.9455)$$
$$F1 = 0.9370 \approx 93.70\ \%$$

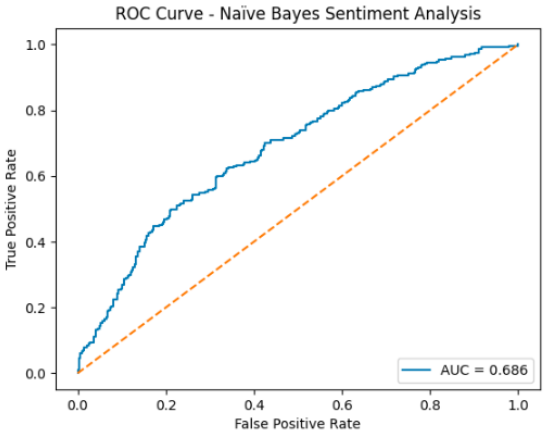

Figure 6. ROC curve result

Figure 6 shows the Overall Performance (AUC = 0.686), with the model having an AUC of approximately 0.686, which means that it performs better than random guessing (0.5) but is still far from a strong classification (≥ 0.8). In probabilistic terms, the model has approximately a 68.6% chance of ranking a positive sentiment higher than a negative one. This is typically classified as "fair" or "moderate" performance, not a strong performance. Furthermore, the ROC curve lies above the diagonal baseline, confirming that the model has learned the meaningful patterns from the data. However, the curve did not approach the upper left corner of the ROC space, which represents optimal performance. This indicates that the model has limitations in achieving high sensitivity without incurring a relatively high false-positive rate. In other words, improvements in correctly identifying positive sentiments are accompanied by an increase in the misclassification of negative instances.

AUC values below 0.80 are generally considered to reflect limited or fair classification performance, particularly in applications that require high reliability. Consequently, although the Naïve Bayes model provides a baseline level of performance for sentiment classification, its predictive capability may not be sufficient for high-stakes decision making. In summary, the ROC analysis indicated that the model achieved moderate classification performance, with clear room for improvement. Enhancing feature representation, incorporating more advanced machine learning algorithms, or applying deep learning-based approaches may improve the discriminative power and overall robustness of the model.

Based on the evaluation results, the Naïve Bayes model achieved an accuracy of 94.17%, indicating that it performed well in classifying sentiment. The precision value of 92.86% shows that most of the predicted positive sentiments were correct, whereas the recall value of 94.55% indicates that the model is highly capable of identifying actual positive sentiments. These results demonstrate that the Naïve Bayes algorithm is effective for analyzing public sentiment toward John Herdman on social media X.

The relatively small number of false negatives (15 cases) indicates that the model rarely failed to detect positive public sentiment. This is important because missing positive opinions could underestimate the public's support for the coach.

Meanwhile, the false-positive count (20 cases) was slightly higher, meaning that some negative opinions were mistakenly interpreted as positive. However, this number remains low compared with the total dataset size.

Corresponding Author: ipan.ardianto@gmail.com
Doi: 10.30996/elsains.v8i1.133496

The obtained F1-score of 93.70% demonstrates that the Naïve Bayes model achieves a strong balance between precision and recall. This result confirms that the model is not only accurate overall but also consistent in correctly identifying both positive and negative sentiments. Therefore, the evaluation based on accuracy, precision, recall, F1-score, and confusion matrix collectively indicates that the proposed sentiment analysis model has a robust classification performance for analyzing public reactions toward the appointment of John Herdman as the head coach of the Indonesian National Football Team.

The novelty of this research is that it provides input to the coaching team, especially the Indonesian national team coach, to conduct analyses based on data and a highly scientific and objective approach. This is necessary considering the position of the national football team coach, which is very vulnerable to being the subject of ridicule when losing to an opponent in a tournament/match. Especially if they lose to a long-standing rival team, such as Malaysia, Vietnam, or Singapore. Providing a new paradigm for local coaches, preparing a team before a match must be carried out in detail, meticulously, and well-planned based on data and the opponent's characteristics.

Analyzing the opponent's gameplay before the match is very important, and should not rely on visual observation and intuition. The analysis includes strengths and weaknesses, which can be expressed statistically, providing a level of objective truth. An evaluation was conducted to ensure that the model created was appropriate and accurate for this research case. The evaluation results determine the next steps, whether the research can continue, or needs to be repeated if it does not align with the initial plan. Overall, the Naïve Bayes algorithm has proven to be an effective method for public sentiment analysis. With an accuracy of 92.45%, this algorithm can accurately identify public sentiment toward coach John Herdman, showing that the Indonesian public has a positive response to his leadership. These results are consistent with other findings that show that Naïve Bayes can handle variations in public sentiment toward important figures and issues.

5 Conclusion

Social media strongly shapes public opinion regarding national team decisions. Sentiment analysis, especially on large-scale social media data, provides a systematic and objective way to understand public attitudes. It moves evaluation away from purely subjective opinions toward measurable insights. Public sentiment toward Herdman is predominantly positive. This positive sentiment reflects satisfaction with leadership expectations and increased trust in team direction. The study also confirms that the Naïve Bayes algorithm achieves an accuracy of up to 94.17% in classifying sentiments (positive vs. negative). This demonstrates that it is reliable for large datasets, can consistently identify patterns in public opinion, and is suitable for real-time monitoring of supporter sentiment. Although effective, the study suggests improvements such as combining Naïve Bayes with other models (e.g., SVM, Random Forest), using deep learning (e.g., LSTM, BERT), incorporating Indonesian slang, emojis, and context, expanding datasets, and including neutral/multimodal analysis. The inclusion of the F1-score and confusion matrix analysis strengthened the evaluation of the proposed model. The Naïve Bayes classifier achieved an F1-score of 93.70%, indicating a strong balance between precision and recall. Furthermore, the confusion matrix shows that the misclassification rates were relatively low, with only 20 false positives and 15 false negatives, confirming the robustness of the model in identifying public sentiment toward John Herdman.

Author Contribution

The first author conceived and designed the research framework and conceptualization, including the formulation of the research problem, objectives, and overall conceptual approach using data mining for sentiment analysis in football. The author was primarily responsible for developing the introduction and research background and aligning the study with current trends in data-driven sports analytics. In addition, writing—review and editing supervised the entire research process to ensure scientific rigor and consistency with the CRISP-DM methodology. This includes analyzing prior studies on sentiment analysis, machine learning approaches, and football analytics and contextualizing the results within the Indonesian football ecosystem. He was also responsible for drafting the Discussion and Conclusion sections, highlighting the implications of the findings, and proposing directions for future research, including the integration of advanced machine learning and deep learning methods.

The second author conducted the data collection and technical implementation of the study. This included extracting Twitter (X) data using Google Colab, performing data preprocessing stages such as cleansing, tokenization, normalization, and labeling, and implementing the Naïve Bayes classification model using Rapidminer. The second author also handled the modeling, testing, and evaluation processes, including accuracy measurement and interpretation of the classification results. All authors contributed to writing, reviewing, and editing the manuscript. All authors have read and approved the final version of this manuscript.

Acknowledgment

We extend our appreciation to the developers and communities behind the tools and platforms utilized in this study, including Google Colab and RapidMiner, which greatly facilitated the processes of data collection, preprocessing, and modeling. We are also grateful to the providers of public data on the social media platform X, whose openly accessible information made this study possible. Special thanks are given to colleagues and peers who provided valuable insights, feedback, and discussions throughout the research process, helping to refine both the methodology and the interpretation of the results. Their constructive input significantly improved the quality of this study.

References

[1] N. Harahap, Al *et al.*, “Evaluasi Kepuasan Penggemar Sepak Bola Terhadap Pemilihan Pelatih Timnas Indonesia Di Media Sosial X Dengan Metode K-Means Clustering,” TechnoCOM, vol. 24, no. 3, pp. 862–876, 2025.

[2] J. She, K. Swart-arries, M. Belal, and S. Wong, “What Sentiment and Fun Facts We Learnt Before FIFA World Cup Qatar 2022 Using Twitter and AI,” arXiv:2306.16049v1 [cs.CL] 28 Jun 2023, Cornell University, Canada, 2023.

[3] L. Ardiantoro, R. M. Akbar, and Y. Nur S., “Tactical Analysis of Indonesian National Football Team Using Fp-Tree Algorithm,” *Proceeding Int. Conf. Sci. Heal. Technol.*, ICOHETECH, UDB, Surakarta, pp. 124–130, 2023, doi: 10.47701/icohetech.v4i1.3381.

[4] M. R. Fahlevi *et al.*, “Sentiment Analysis of Public Opinion on PSSI Naturalization Program Based on Social Media Using the Naive Bayes Algorithm,”JEECS, vol. 10, no. 2, pp. 160–167, 2025.

[5] H. Mahmud, H. Mahmud, M. Rifat, and A. Rashid, “Enhancing Sentiment Analysis in Bengali Texts : A Hybrid Approach Using Lexicon-Based Algorithm and Pretrained Language Model Bangla-BERT”, arXiv:2411.19584v2 [cs.LG] 23 Apr 2025, Cornell University, 2025.

[6] L. Ardiantoro, S. Zahara, and N. Sunarmi, “Pemanfaatan

Corresponding Author: ipan.ardianto@gmail.com

Doi : 10.30996/elsains.v8i1.133496

Knowledge Data Discovery (KDD) pada Pola Permainan Atlet Bulutangkis," Online, 2019. [Online]. Available: http://jurnal.yudharta.ac.id/v2/index.php/EXPLORE-IT/

[7] W. B. Zulfikar, A. R. Atmadja, and S. F. Pratama, "Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes,", http://journal.unnes.ac.id/nju/index.php/sji, Scientific Journal of Informatics, vol. 10, no. 1, pp. 25–34, 2023 e-ISSN 2460-0040, doi: 10.15294/sji.v10i1.39952.

[8] S. Mujilahwati, N. Z. M. Safar, and C. Supriyanto, "Segmentation Word to Improve Performance Sentiment Analysis for Indonesian Language, 15, no. 02, pp. 145–154, 2024.

[9] C. Juliane, "Public Response to the Constitutional Court ' s Decision on Indonesia ' s 2024 Elections," Research Square, pp. 1–15, 2024.

[10] A. A. Firdaus, A. Yudhana, and I. Riadi, "Public opinion analysis of presidential candidate using naïve bayes method," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, http://kinetik.umm.ac.id, ISSN: 2503-2267, vol. 4, no. 2, 2023.

[11] L. W. Astuti and Y. Sari, "Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data," International Journal of Advanced Computer Science and Applications, vol. 14, no. 10, pp. 498–504, 2023.

[12] M. Irfan, S. Basuki and Y. Azhar, "Impact of Feature Selection and Data Augmentation for Pregnancy Risk Detection in Indonesia," www.umm.ac.id, vol. 12, no. 6, pp. 2266–2273, 2022.

[13] F. K. Basri, M. Afdal, A. Angraini, and N. E. Rozanda, "Analisis Sentimen Masyarakat Terhadap Kebocoran Pusat Data Nasional Sementara Menggunakan Algoritma Random Forest dan Support Vector Machine," vol. 7, no. 2, pp. 960–971, 2025, doi: 10.47065/bits.v7i2.7473.

[14] A. Romadhony, "Sentiment Analysis on a Large Indonesian Product Review Dataset," vol. 10, no. 1, pp. 167–178, 2024.

[15] J. Su, R. Y. K. Lau, J. Yu, D. Chi, T. Ng, and W. Jiang, "A multi-modal data fusion approach for evaluating the impact of extreme public sentiments on corporate credit ratings," *Complex Intell. Syst.*, vol. 11, no. 10, pp. 1–16, 2025, doi: 10.1007/s40747-025-02067-5.

[16] Z. A. Anwer, "Assessing Institutional Performance using Machine Learning on Arabic Facebook Comments," *Eng. Technol. Appl. Sci. Res. , vol. 14, no. 4, p. 16025, Aug. 2024, doi 10.48084/etasr.8079.*, vol. 14, no. 4, pp. 16025–16031, 2024.

[17] A. M. Joshy, J. N. Thomas, A. M. Jose, E. Talit, and S. George, "Tweet Pulse : Interactive Sentiment Dashboard with Advanced Analytics," vol. 15, no. 6, pp. 32–39, 2025, doi: 10.48047/ijiee.2025.15.6.6.

[18] M. Gupta and A. Kaushik, "Unveiling Public Perceptions : Machine Learning-Based Sentiment Analysis of COVID-19 Vaccines in India".

[19] A. P. Wibawa, D. E. Cahyani, D. D. Prasetya, and L. Gumilar, "Detecting emotions using a combination of bidirectional encoder representations from transformers embedding and bidirectional long short-term memory," vol. 13, no. 6, pp. 7137–7146, 2023, doi: 10.11591/ijece.v13i6.pp7137-7146.

[20] K. Chaudhary, M. Alam, *Big Data Analytics; Applications in Business and Marketing*, vol. 1, no. 1. 2022. Available at: https://www.routledge.com/Big-Data-Analytics-Applications-in-Business-and-Marketing/Chaudhary-Alam/p/book/9781032187662

[21] B. P. Aji, C. Sri, and K. Aditya, "Klasifikasi Sentimen Ulasan Produk pada Platform E-Commerce di Indonesia dengan Menggunakan Model Pre-Trained IndoBERT," vol. 6, no. 4, 2025, doi: 10.47065/bits.v6i4.6968.

[22] C. Series, "The accuracy of transfer learning using neural network method for sentiment analysis problem on Indonesian tweets," 2022, doi: 10.1088/1742-6596/1725/1/012015.

[23] S. Riyadi, M. D. Mubarok, C. Damarjati, and A. J. Ishak, "Improving Sentiment Analysis Accuracy Using CRNN on Imbalanced Data : a Case Study of Indonesian National Football Coach," vol. 12, no. 2, pp. 159–167, 2024.

[24] H. Han, M. Asif, E. M. Awwad, N. Sarhan, Y. Y. Ghadi, and B. Xu, "Innovative deep learning techniques for monitoring aggressive behavior in social media posts," 2024, doi: 10.1186/s13677-023-00577-6.

[25] A. Glenn, A. Glenn, and B. Cox, "Emotion classification of Indonesian Tweets using Bidirectional LSTM," *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9567–9578, 2023, doi: 10.1007/s00521-022-08186-1.

[26] A. Albladi, K. Uddin, M. Islam and C. Seals, "TWSSenti : A Novel Hybrid Framework for Topic-Wise Sentiment Analysis on Social Media Using Transformer Models," 2014.

[27] M. Mishra, "AI and Analitic in Sport: Leveraging BERTopics to map the past and chart the future.", arXrv Cornell University, pp. 1–32, Oct 2025, doi: 10.48550/arxiv.2510.15487.

Corresponding Author: ipan.ardianto@gmail.com
Doi: 10.30996/elsains.v8i1.133496