Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13275
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analyzing Public Sentiment Regarding the Qatar 2023 World Cup Debate Using TF-IDF and K-Nearest Neighbor Weighting

Sayyid Muh. Raziq Olajuwon[1)], Kusrini[2)] , Kusnawi[3)]
[1,2,3)] Informatics Engineering Study Program, Amikom University Yogyakarta
[1)]olajuwonda@students.amikom.ac.id, [2)]kusrini@amikom.ac.id,[3)]Khusnawi@amikom.ac.id

Submitted : Jan 20, 2020 | **Accepted** : April 20, 2020 | **Published** : July 31, 2020

Abstract: This research aims to uncover the sentiment of Twitter users regarding the polemics surrounding the 2023 Qatar World Cup using a text-based sentiment analysis approach. The research methodology involves collecting data from Twitter posts, encompassing discussions, opinions, and responses related to the Qatar World Cup 2023. The TF-IDF weighting is applied to identify significant keywords in each post, while the K-Nearest Neighbor algorithm is employed to classify sentiments as positive, negative, or neutral. The findings reveal a comprehensive picture of how the public perceives the Qatar World Cup 2023 on the Twitter platform. The results not only cover positive and negative aspects of online discussions but also identify trends and patterns of sentiment that emerge during specific periods.The application of these methods provides valuable insights into understanding the dynamics of public opinion related to international sports events through the lens of social media. The results of the analysis demonstrate that a majority of Twitter users express positive sentiments towards the Qatar World Cup 2023, highlighting excitement and anticipation. However, some negative sentiments also arise, primarily related to controversies and concerns about the event. The research further identifies temporal variations in sentiment, reflecting changing public perceptions over time.This research contributes to the development of sentiment analysis methods by using a combination of TF-IDF weighting and the K-Nearest Neighbor algorithm to delve into Twitter users' perspectives. Consequently, the findings have practical applicability for further research and implementation in managing the social impact and public perception of major sporting events like the World Cup. .

Keywords: K-Nearest Neighbor, Sentiment Analysis, TF-IDF, Qatar World Cup

INTRODUCTION

A topic that is currently being discussed on the Twitter platform is the 2022 FIFA World Cup. On December 2, 2010, in a surprise move, Qatar won the bid to host the 2022 FIFA World Cup, making Qatar the first Middle Eastern country to achieve the honor. Qatar was chosen as the host after beating Australia, Japan, South Korea, and the United States who were also vying to host the 2022 World Cup. The selection of Qatar as the host of the 2022 World Cup can lead to various responses from various circles around the world. Social media such as Twitter is one of the platforms to express responses or opinions. This can be used as a basis for conducting sentiment analysis on Qatar as the host of the 2022 World Cup (Dewi & Arianto, 2023).

When dealing with the complexity of issues related to the Qatar 2023 World Cup, it is important to understand how people are responding to it on social media. Sentiment analysis is a relevant approach to identify the views, dominant thoughts, and key debates developing in this context. A sentiment analysis of Twitter users towards the Qatar 2023 World Cup by utilizing the TF-IDF (Term Frequency-Inverse Document Frequency) clustering technique and a comparison of Support Vector Machine, K-Nearest Neighbor, and Random Forest methods will be able to identify people's sentiment within the social media sphere for the Qatar 2023 World Cup topic.

To carry out the analysis, Google Colab is used, a cloud-based platform that makes it possible to run sentiment analysis in the Python programming language. In this research, the Google Colab platform is used, which utilizes the Python programming language. The use of Python in analyzing sentiment becomes more efficient thanks to the various libraries that support it (Maulana et al., 2023). This approach will provide deep insights into various

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

aspects related to the tournament, including people's responses to Qatar's selection as the host, human rights issues, economic impact, and other debates.

This analysis will be run using Python, which is a powerful programming language that is well-suited for text processing and data analysis. Thus, the background of this problem comprehensively reflects the urgency and relevance of this research in revealing people's sentiments towards the Qatar 2023 World Cup through social media analysis, which can provide valuable insights in the ever-changing global context.

In summary, the selection of Qatar as the host of the 2022 FIFA World Cup has sparked significant global interest and discussions. The anticipation and concerns surrounding this decision, along with the multifaceted issues associated with the Qatar 2023 World Cup, underscore the need for a comprehensive analysis of public sentiments. By employing advanced sentiment analysis techniques and leveraging Python-based tools like Google Colab, this research aims to uncover and understand the nuanced perspectives expressed by Twitter users. The findings are expected to contribute valuable insights into the complex landscape of opinions surrounding the Qatar 2023 World Cup on social media, shedding light on various dimensions such as public perception, controversies, and emerging trends.

LITERATURE REVIEW

Related Research

Research with the title "Twitter User Sentiment Analysis on PPKM Extension Using the K-Nearest Neighbor Method". This journal evaluates the sentiment of Twitter users regarding the extension of the PPKM Policy (Enforcement of Restrictions on Community Activities) which often gets negative responses on social media. This research uses the K-Nearest Neighbor (K-NN) algorithm and tweet data with the keyword "PPKM" collected during a certain period. The model training results achieved an accuracy score of 69.5%, which shows the model's ability to classify Twitter user sentiment towards PPKM (Asro'i & Februariyanti, 2022).Next "Sentiment Analysis on Twitter Social Media Towards Student Academic Information System Services at Universitas Brawijaya with the K-Nearest Neighbor Method". This research focuses on analyzing the sentiment of Twitter users towards the Brawijaya University Student Academic Information System Service (SIAM UB). The K-Nearest Neighbor (K-NN) method is used to classify tweets into positive or negative sentiment classes. The results achieved the best accuracy of 86% with the use of k value = 3 and 100% features (Dharmawan et al., 2020).Then "Sentiment Analysis Using K-Nearest Neighbor Towards New Normal of the Covid-19 Period in Indonesia". This research focuses on analyzing the sentiment towards the "New Normal" policy during the COVID-19 pandemic in Indonesia, which has become a hot topic on Twitter social media. Data from 1000 tweets are used for model training using K-Nearest Neighbor (K-NN). Classification results with KNN (k = 1) achieved high accuracy, namely 100% on the training set, 92.60% on 10-fold cross-validation,
and 94.50% on 80% percentage split (Furqan et al., 2022).Next is "Implementation of K-Nearest Neighbor (K-NN) Algorithm for Public Sentiment Analysis of Online Learning". This research applies the K-Nearest Neighbor (K-NN) algorithm in sentiment analysis of Twitter users related to online learning. Indonesian tweet data from February to September 2020 was used in this study. The highest accuracy results were obtained at k = 10 with 84.65% accuracy. The research also revealed that public opinion tends to be positive towards online learning (Isnain et al., 2021).Then there is the research "Expert System for Diagnosing Autism in Android-Based Children". This journal discusses the implementation of an expert system to diagnose autism in Android-based children. This expert system helps in the process of identifying autism quickly using data from psychologists and special needs teachers. The test results show that this system can provide an overview of autistic children and appropriate therapy methods (Nurhakim et al., 2017).Finally, the research "Comparing Sentiment Analysis of Indonesian Presidential Election 2019 with Support Vector Machine and K-Nearest Neighbor Algorithm". This research compares the effectiveness of the Support Vector Machine (SVM) algorithm with the K-Nearest Neighbor (K-NN) algorithm in predicting the results of the 2019 Indonesian presidential election based on sentiment analysis on social media (Twitter). The results show that SVM has higher accuracy than K-NN, with an average accuracy value of 69.27% for SVM and 61.3% for K-NN. Predictions based on positive sentiment show different results for both presidential candidates (Kristiyanti et al., 2019).

Theoretical Foundation

Data Analysis

Data analysis is the systematic process of finding and organizing the transcripts, interviews, field notes and other materials you have collected. The aim is to improve your personal understanding of the information and enable you to present your findings to others (Rijali, 2018).

Sentiment Analysis

Sentiment analysis is one of the sub-disciplines in text mining research that deals with broader aspects such as data processing in certain activities (Dedi Darwis et al., 2020).

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Twitter

Currently, the social media application Twitter is very popular among internet users. Based on research conducted in November 2019, there are around 78 million Twitter users in Indonesia (Normawati & Prayogi, 2021).

TF-IDF Weighting

TF-IDF is one of the methods used in text processing to assign importance to words in a document. The function of TF-IDF is to recognize words that have the highest significance in a document or group of documents (Wati et al., 2023).

Support Vector Machine Method

A commonly used classification algorithm for sentiment analysis is the Support Vector Machine (SVM) method. The main goal of SVM is to identify the optimal hyperplane. The optimal hyperplane is the hyperplane that is in the middle of the two classes so that it has the furthest distance to the outermost data of each class. SVM tries to maximize the margin between the two classes by finding the right hyperplane (Novantika, 2022).

K-Nearest Neighbor Method

The KNN algorithm is a technique used to classify data by considering the closest distance between the data to be classified and the existing data. Determining the optimal K value for this algorithm depends on the data being processed. A large K value can reduce the influence of noise on the classification process, but can also make the boundaries between classes less firm (Homepage et al., 2021).

Random Forest Method

The Random Forest Classifier is a randomized ensemble based on a decision tree. Random Forest model construction involves several key steps. First, an n-tree bootstrap sampling of the data is performed. Next, for each bootstrap data set, the tree is grown by randomly selecting variables to separate the nodes in the tree. The tree is expanded so that each terminal node has a sufficient number of cases. The next stage involves combining the information from all trees in the ensemble to predict new data, generally by using majority voting in classification. Finally, the out-of-bag (OOB) error rate is calculated by utilizing data not included in the bootstrap sample (Normah et al., 2022).

Python

Python is a high-level programming language that is interpreter, interactive, object-oriented, and can run on almost all types of platforms such as Linux, Windows, Mac, and other systems. Python is also known as a high-level programming language that is relatively easy to learn because it has a clear and elegant syntax. In addition, Python also utilizes various modules with high-level data structures that are efficient, ready to use, and simplify the application development process (Ratna, 2020).

Google Colab

Google Colab is a research project from Google created to help spread education and research in the field of machine learning. It is a Jupyter notebook environment that can be used without any setup and runs entirely in the cloud (Ray et al., 2021).

METHOD

This research is quantitative in nature, utilizing numerical data to measure, analyze, and test relationships between specific variables. The focus is on collecting numerical data related to Twitter users' sentiments toward the 2023 Qatar World Cup and analyzing it using the K-Nearest Neighbor (K-NN) algorithm within a quantitative approach.With a descriptive nature, this research aims to depict phenomena without intervention or alteration. Data collection is performed through the web scraping technique, which is a method for automatically extracting information from web pages. In the context of this research, web scraping will be employed to retrieve tweets with the keyword "2023 Qatar World Cup" during a specific period from the Twitter platform.

The web scraping process begins by designing a computer program or script capable of navigating Twitter pages, extracting information from relevant HTML elements, and storing it in a dataset. It is essential to emphasize that the use of this technique must adhere to Twitter's data usage policies and research ethics principles.After successfully gathering the data, the next step involves implementing sentiment analysis using the K-NN algorithm. The collected data includes tweet text, posting dates, retweet counts, like counts, and user information. No survey questionnaire is required for this research. By integrating web scraping techniques with sentiment analysis, this research aims to provide in-depth insights into Twitter users' perspectives and opinions regarding the 2023 Qatar World Cup on a quantitative scale.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13275
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

Research Stages

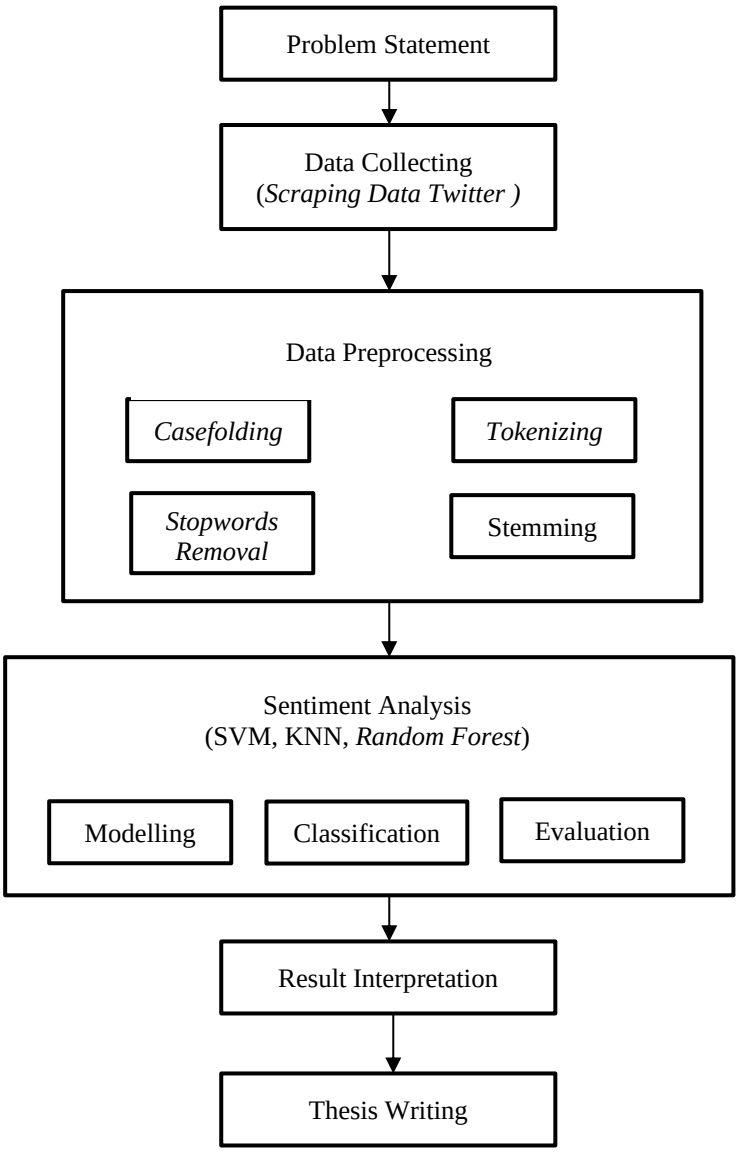

Fig. 1 Research Stages

This research will be divided into several stages as follows:

Problem Formulation

This problem formulation integrates findings from literature reviews related to sentiment analysis of the 2023 Qatar World Cup on Twitter. Considering public responses to Qatar's selection as the host, potential controversies, economic impacts, and human rights issues, the research problem is articulated clearly and relevantly.The sentiment analysis method employs TF-IDF clustering and a comparative study of Support Vector Machine, K-Nearest Neighbor, and Random Forest methods. The goal is to identify dominant perspectives, key debates, and sentiment patterns on the Twitter platform concerning the 2023 Qatar World Cup.Through this problem formulation, the research aims to make a significant contribution to understanding public opinions through social media, addressing knowledge gaps, and providing a solid foundation to achieve research objectives.

Data Collection

After formulating the background of the problem, the next step is to collect data as the research source. In this case, the required data consists of Tweets related to the 2022 Qatar World Cup.The data collection method employed in this research involves web scraping from the Twitter platform. Web scraping is a relevant approach for accessing and collecting large-scale Tweet data. This process entails extracting information from Twitter web

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

pages using specific techniques.The gathered data includes various aspects, including but not limited to public responses to the 2022 Qatar World Cup, opinions, retweets, and relevant hashtags. It's important to note that web scraping will be conducted while ensuring compliance with Twitter's usage policies and data collection ethics. All necessary steps will be taken to ensure that the data is obtained legally and in accordance with applicable rules.By implementing this data collection method, it is expected that the obtained dataset will provide a solid foundation for analyzing Twitter users' sentiments towards the 2022 Qatar World Cup, aligning with the previously formulated research focus.

Prepocessing Data

After successfully collecting Tweets related to the 2022 Qatar World Cup, the next crucial step involves preprocessing the data for a more in-depth sentiment analysis. This process entails a series of detailed steps outlined as follows: Firstly, we apply case folding, converting all text in the dataset to lowercase. The aim of this step is to achieve consistency in the analysis, given the frequent use of varying capitalization on the Twitter platform. The second step is tokenizing, where text is broken down into tokens or individual words. This allows for a more detailed analysis at the word level, facilitating the identification of sentiment in a more specific manner. The subsequent stage involves stopwords removal. Common stopwords such as "and," "the," and "is" are eliminated to enhance the precision of the analysis by focusing on words that carry specific meaning related to sentiment. The final step is stemming, where words are transformed into their base form. This aids in reducing word variation, making it easier for comparison and analysis in subsequent processes.By implementing these steps, the Tweet data has been effectively prepared for a more in-depth sentiment analysis concerning the 2022 Qatar World Cup.

Sentiment Analysis

After successfully processing the data, this study proceeds to sentiment analysis, encompassing modeling, classification, and evaluation.In the modeling phase, we employ TF-IDF clustering and compare it with SVM, KNN, and Random Forest methods to identify and categorize sentiment patterns within the dataset of Tweets related to the 2022 Qatar World Cup.The classification process involves applying SVM, KNN, and Random Forest algorithms to categorize each Tweet into positive, negative, or neutral sentiments. This step aims to provide a deeper insight into Twitter users' perspectives on the 2022 Qatar World Cup.The final step is evaluation, utilizing standard metrics such as accuracy, precision, recall, and F1-score to assess the model's performance. This evaluation is crucial to validate the accuracy of the sentiment analysis results. By integrating modeling, classification, and evaluation, this research aims to present a comprehensive understanding of Twitter users' sentiments regarding the 2022 Qatar World Cup, contributing valuable insights into public opinion through social media.

Interpretation of Results

In interpreting the results, the sentiment analysis method with the highest accuracy will be considered the most suitable approach. Subsequently, this high-accuracy method will undergo a detailed interpretation as the primary model. This step is taken to ensure that the obtained sentiment analysis results achieve optimal accuracy.After identifying the method with the highest accuracy, the next step is to retest the model using this best-performing method, focusing specifically on opinions related to the "2022 Qatar World Cup." This process aims to obtain sentiment analysis results that are more specific and relevant to the specified keyword. Additionally, a word cloud will be generated to reflect the most prevalent opinions within the research dataset. This word cloud will provide a clear visualization of the words that dominate in expressing sentiment among Twitter users regarding the 2022 Qatar World Cup. Thus, this interpretation of results not only measures the accuracy of the model but also offers a deeper understanding of the prevailing opinions in online conversations related to this sporting event.

Thesis Writing

After successfully interpreting the sentiment analysis results with precision, the next step is to translate the obtained information into a comprehensive research thesis report. This thesis writing process will encapsulate the findings, interpretations, and conclusions derived from the sentiment analysis of Twitter user conversations regarding the 2022 Qatar World Cup.The thesis report will intricately detail the methodology employed in data collection and processing, present the findings from the sentiment analysis, and describe the interpretation methods applied to the model with the highest accuracy. In this section, particular emphasis will be placed on a deeper understanding of public opinions, especially those related to the keyword "2022 Qatar World Cup." Furthermore, the thesis writing will involve the development and elaboration of the word cloud analysis results that reflect the most dominant opinions. By detailing these findings, the thesis aims to make a valuable

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13275
e-ISSN : 2541-2019
p-ISSN : 2541-044X

contribution to the understanding of the dynamics of public opinion through social media, particularly concerning a major international sporting event like the Qatar World Cup.

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" |

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

Fig. 2 Twitter Data Crawling Script Using Twiteharvest

From the illustration above, the utilization of the pandas library and TwiteHarvest from a third-party provider, namely JustAnotherArchivist, is evident, which is taken from the GitHub repository. The use of the pandas library in this research aims to present data in the form of tables or dataframes. Subsequently, the necessary modules are imported, namely Pandas, TwiteHarvest (a replacement for Snscrape for Twitter data retrieval), and itertools. The itertools module is used to perform the division of data to be crawled.

	id_str	full_text
2	1,68E+18	Piala dunia U-17 di Stadion Si Jalak Harupat, kemarin diinspeksi Pak Presiden @jokowi bersama ketua PSSI Pak @erickthohir, Menpora Pak Ario Bimo dan Menteri PUPR
3	1,68E+18	PSSI bersama klub sepak bola Indonesia mencari talenta Garuda Muda di 12 kota. Sudah siap menjadi saksi lahirnya talenta muda Indonesia menuju Piala Dunia U-17? ð
4	1,68E+18	Tepat sembilan tahun lalu, Argentina gagal di final Piala Dunia 2014 di Brasil ðŸ‡¦ðŸ‡·ðŸ†šðŸ‡§ðŸ‡· Kegagalan tersebut yang membuat trofi Piala Dunia di Qatar 2022 le
5	1,68E+18	ðŸ“† Berikut jadwal seleksi Tim U-17 Indonesia di 12 kota. Sudah siap menjadi saksi lahirnya talenta muda Indonesia menuju Piala Dunia U-17? ðŸ‘€ #KitaGaruda #Garuc
6	1,68E+18	Karena Anies piala dunia U-17 bisa diselenggarakan di JIS &amp karena Ganjar piala dunia U-20 Gagal digelar.. Ini fakta, silahkan rakyat menilai!! https://t.co/Cp7bVJuX
7	1,68E+18	Pagi tadi di Stadion Si Jalak Harupat, PSSI memulai seleksi pemain untuk Timnas U-17 yang diproyeksikan tampil di Piala Dunia akhir tahun ini. Setelah menggelar seleks
8	1,68E+18	Sekarang, tinggal pemeriksaan ulang oleh FIFA yang memiliki kewenangan untuk menentukan kelayakan stadion sebagai tempat penyelenggaraan Piala Dunia U-17.
9	1,68E+18	Berkunjung ke Stadion Si Jalak Harupat di Kab. Bandung, yang akan jadi salah satu venue pertandingan sepakbola Piala Dunia U-17. Sarana dan prasarana stadion ini sud
10	1,68E+18	Presiden @jokowi meninjau Stadion Si Jalak Harupat yang berada di Kabupaten Bandung. Dalam kesempatan tersebut, Presiden memastikan bahwa Stadion Si Jalak Har
11	1,68E+18	ðŸ‡®ðŸ‡©—Ð¸Bakal calon presiden (Bacapres) Anies Baswedan menanggapi santai soal polemik JIS yang dibangun di eranya saat menjabat sebagai Gubernur DKI Jakarta, y
12	1,68E+18	13 tahun yang lalu, Spanyol menjadi juara Piala Dunia 2010 ðŸ‡ªðŸ‡¸ðŸ‡¿ðŸ‡¦ðŸ‡¿ Waka Waka ðŸŽ¶ https://t.co/Y4kQ32ktcF
13	1,68E+18	Al-Ittihad ingin mengumpulkan gelandang pemenang Piala Dunia 2018 ðŸ˜¬ #Alittihad #Transfer https://t.co/xcHbXOJstJ
14	1,68E+18	Pernyataan Resmi Bos Buro Happold Menunjukkan JIS yang Dibangun Anies Sangat Layak Jadi Venue Piala Dunia https://t.co/ySTHKDthzq
15	1,68E+18	Oiii kita di Kibulin ! Piala Dunia U20 itu batal Seolah-olah karna Menolak kedatangan Timnas Israel, " Padahal ada Faktor KETIDAKSIAPAN Infrastruktur " (sumbe
16	1,68E+18	âœ... RESMI: 34 pemain yang dipanggil kuch Bima Sakti untuk menjalani TC jelang persiapan Piala Dunia U-17. Terdapat 6 pemain keturunan yang dipanggil, seperti: Â· W
17	1,68E+18	Pernyataan resmi manager Buro Happlod for Indonesia, JIS dibangun berstandar FIFA &amp sangat layak diselenggarakan piala dunia.. Buzzer auto mingkem!! https://
18	1,68E+18	Makin panjang dan melebar ini pasti gorang-goreng terkait JIS. Udah lah, tutup aja kalau emang tuh Stadion digorat-geret kanan-kiri oleh Cebong, Kampret, Kadrun, bab
19	1,68E+18	Video pengumuman skuat Timnas Wanita Filipina yang akan tampil di Piala Dunia 2023 Australia - Selandia Baru. Sejarah bagi Filipina pertama kali tampil di event sepak
20	1,68E+18	Adidas terkenal dengan tiga garisnyanya. Terkadang mereka menyelipkan ikon tersebut di jersey tim. Seperti yang terlihat di jersey Belanda pada Piala Dunia 1974. Nan
21	1,68E+18	Batal jadi Tuan Rumah World Beach Games karena dana tidak turun tapi berhasil menerima penunjukkan secara mendadak jadi Tuan Rumah Piala Dunia U-16 âœŒï¸

Fig. 3 World Cup Dataset

```
[ ] !pip install nltk
    !pip install Sastrawi

    import pandas as pd
    import re

[ ] data = pd.read_csv("pialadunia.csv", sep=";")

[ ] data.head()
```

	id_str	full_text	quote_count	reply_count	retweet_count	favorite_count	Unnamed: 6	user_id_str	conversation_id_str	username
0	1,67945E+18	Piala dunia U-17 di Stadion Si Jalak Harupat, ...	12	55	38	335	NaN	80323736	1,67945E+18	ridwankamil
1	1,67936E+18	PSSI bersama klub sepak bola Indonesia mencari...	9	40	30	333	NaN	2363027508	1,67936E+18	PSSI
2	1,67934E+18	Tepat sembilan tahun lalu, Argentina gagal di ...	0	2	10	166	NaN	27197792	1,67934E+18	GOAL_ID
3	1,67911E+18	📆 Berikut jadwal seleksi Tim U-17 Indonesia di...	8	19	36	303	NaN	2363027508	1,67911E+18	PSSI
4	1,67908E+18	Karena Anies piala dunia U-17 bisa diselenggar...	98	1228	862	3151	NaN	1,00683E+18	1,67908E+18	ekowboy2

```
[ ] data.shape
    (113, 11)
```

Fig. 4 Input The Dataset Into The Preprocessing File

The picture above shows the utilization of two libraries, namely NLTK and Sastrawi, which function as data dictionaries. However, it should be noted that these two libraries focus on different languages. NLTK is used to process data dictionaries in English, while Sastrawi is specifically designed for data dictionaries in Indonesian. Additionally, in the displayed code, the 'data' variable is used to read the dataset.csv file and display the imported data.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13275
e-ISSN : 2541-2019
p-ISSN : 2541-044X

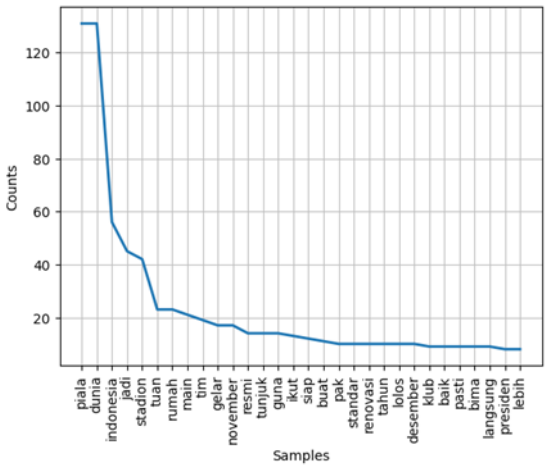

Fig. 5 Visualization Graph of Frequency Data

text_clean	normal	token	normal_1	stop	stemmed	normal_2	teks	text
piala dunia di adion si jalak harupat kemari...	piala dunia di stadion si jalak harupat kemari...	['piala', 'dunia', 'di', 'stadion', 'si', 'jal...	['piala', 'dunia', 'di', 'stadion', 'si', 'jal...	['piala', 'dunia', 'stadion', 'si', 'jalak', '...	['piala', 'dunia', 'stadion', 'si', 'jalak', '...	['piala', 'dunia', 'stadion', 'si', 'jalak', '...	piala dunia stadion si jalak kemarin inspeksi ...	piala dunia stadion si jalak kemarin inspeksi ...
pssi bersama b sepak bola indonesia mencari...	pssi bersama klub sepak bola indonesia mencari...	['pssi', 'bersama', 'klub', 'sepak', 'bola', '...	['pssi', 'bersama', 'klub', 'sepak', 'bola', '...	['pssi', 'bersama', 'klub', 'sepak', 'bola', '...	['pssi', 'sama', 'klub', 'sepak', 'bola', 'ind...	['sama', 'klub', 'sepak', 'bola', 'indonesia',...	sama klub sepak bola indonesia cari talenta ga...	sama klub sepak bola indonesia cari talenta ga...

Fig. 6 Clean Data Results

```
[ ] # Pembuatan Kamus Lexicon
    lexicon_positive = {}
    with open('positive.tsv', 'r') as tsvfile:
        reader = csv.reader(tsvfile, delimiter='\t')
        for row in reader:
            if len(row) > 1 and row[1].lstrip('-').isdigit():
                lexicon_positive[row[0]] = int(row[1])

    lexicon_negative = {}
    with open('negative.tsv', 'r') as tsvfile:
        reader = csv.reader(tsvfile, delimiter='\t')
        for row in reader:
            if len(row) > 1 and row[1].lstrip('-').isdigit():
                lexicon_negative[row[0]] = int(row[1])

    def sentiment_analysis_lexicon_indonesia(text):
        score = 0
        words = text.split()
        for word in words:
            if word in lexicon_positive:
                score += lexicon_positive[word]
            if word in lexicon_negative:
                score -= abs(lexicon_negative[word])  # Mengurangkan bobot negatif
        return score

[ ] df['teks'] = df['stemmed'].str.replace('\'', '')
    df['teks'] = df['teks'].apply(lambda x: x.replace('[','').replace(']',''))
    df['teks'] = df['teks'].str.replace(',', '')

[ ] df['compound'] = df['teks'].apply(sentiment_analysis_lexicon_indonesia)

    df['polarity'] = df['compound'].apply(lambda score: 'positive' if score > 0 else ('negative' if score < 0 else 'neutral'))

    print(df['polarity'].value_counts())
```

Fig. 7 Labeling Script

The picture above shows the labeling script, this process helps in classifying whether a piece of data has a positive, negative, or neutral sentiment based on the overall weight generated. In this way, sentiment analysis can provide a clearer understanding of the sentiments contained in the dataset based on the lexicon dictionary and weighting that has been applied.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13275
e-ISSN : 2541-2019
p-ISSN : 2541-044X

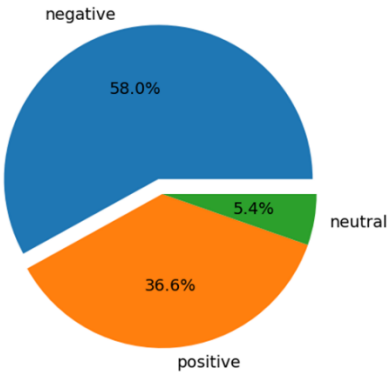

Fig. 8 Visualization of Pie Chart for Labeling Results

The above pie chart visualizes the distribution of data in the dataset related to opinions about the 2023 Qatar World Cup. According to the chart, data labeled as "positive" dominates approximately 36.6% of the total dataset. On the other hand, data labeled as "negative" reaches 58%, while data labeled as "neutral" contributes around 5.4%. With these results, it can be concluded that the majority of opinions in the dataset tend to be negative regarding the controversy of the 2023 Qatar World Cup.

```
[10] from sklearn.model_selection import train_test_split
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
     from sklearn.feature_extraction.text import TfidfVectorizer

     def classify_and_evaluate(X_train, y_train, X_test, y_test, n_neighbors=5):
         # Initialize the KNN classifier
         model = KNeighborsClassifier(n_neighbors=n_neighbors)

         # Train the model
         model.fit(X_train, y_train)

         # Predict on the test set
         y_pred = model.predict(X_test)

         # Print classification report, confusion matrix, and accuracy score
         classification_rep = classification_report(y_test, y_pred, zero_division=0)
         print("============================================================")
         print("Classification Report:\n", classification_rep)
         print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
         print("Accuracy Score:\n", accuracy_score(y_test, y_pred))

         return classification_rep

[19] # Split data into training and test set
     X_train, X_test, y_train, y_test = train_test_split(df['teks'], df['polarity'], test_size=0.3, random_state=42)

     # Convert training and test set to TF-IDF representation with Laplace Smoothing
     vectorizer = TfidfVectorizer(smooth_idf=True, sublinear_tf=False)
     X_train_tfidf = vectorizer.fit_transform(X_train)
     X_test_tfidf = vectorizer.transform(X_test)
```

Fig. 9 Script for Data Splitting and the Utilization of K-Nearest Neighbors (KNN) Method

```
============================================================
Classification Report:
              precision    recall  f1-score   support

    negative       0.50      0.81      0.62        16
     neutral       0.00      0.00      0.00         2
    positive       0.62      0.31      0.42        16

    accuracy                           0.53        34
   macro avg       0.38      0.38      0.35        34
weighted avg       0.53      0.53      0.49        34

Confusion Matrix:
 [[13  0  3]
 [ 2  0  0]
 [11  0  5]]
Accuracy Score:
 0.5294117647058824
```

Fig. 10 Output Classification Report

The picture above is a script for implementing data splitting and the Naive Bayes method. First, the neutral label is dropped or removed, then the data is split with a division of 70% for training data and 30% for test data.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Subsequently, the K-Nearest Neighbors (KNN) method is used for analysis, followed by the output classification report.

Results :

This study explores Twitter users' sentiments regarding the 2023 Qatar World Cup by employing sentiment analysis methods using TF-IDF weighting and classification algorithms, including Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest (RF) on Google Colab. The evaluation results of the models indicate that each algorithm provides an overview of its ability to classify sentiments. Although specific accuracy and F1 score figures are not disclosed at this stage, qualitative analysis offers in-depth insights into the relative performance of each model. The study's limitations involve focusing on Twitter data with the keyword or hashtag "2023 Qatar World Cup." The selection of three sentiment categories (positive, negative, and neutral) facilitates interpretation and analysis.

Detailed Results and Analysis Process:

In this study, sentiment results are obtained through the implementation of TF-IDF weighting. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method that measures the relevance of specific words in a document compared to the entire document collection. The TF-IDF process assigns scores to words based on how often they appear in a specific document (Term Frequency) and how unique or uncommon the words are across the document collection (Inverse Document Frequency). Subsequently, the TF-IDF results are used as features in the K-Nearest Neighbor (K-NN) algorithm for sentiment classification. K-NN is a classification algorithm that utilizes the proximity between data points. In this context, data points closest to the test data (tweets to be classified) will have a greater influence in determining sentiment. In other words, if a tweet contains words similar to previously classified tweets, it is likely to have a similar sentiment. The derivative processes from TF-IDF formulas and their relationship with K-NN are crucial in producing more accurate sentiment interpretations in the context of Twitter conversations about the 2023 Qatar World Cup.

DISCUSSIONS

In the interpretative phase, a meticulous examination was undertaken by applying keywords specifically related to the 2023 Qatar World Cup, culminating in the creation of a word cloud. This visually immersive representation offers a nuanced and detailed context, enabling a deeper understanding of the prevalent sentiments resonating within the Twitter community.

Contributions to Sentiment Understanding:

This study stands as a substantial contribution to unraveling the intricate tapestry of Twitter users' sentiments regarding the 2023 Qatar World Cup. While the detailed breakdown of the model evaluation results is not explicitly provided at this juncture, the insights garnered offer a comprehensive understanding of the divergences in the classification capabilities exhibited by each algorithm.

Word Cloud as a Visual Synthesis:

The strategic utilization of a word cloud emerges as a particularly potent visual instrument in encapsulating the essence of opinions and sentiments prominent in online conversations. This graphical representation not only succinctly summarizes the prevailing sentiments but also serves as a visually compelling tool, providing a comprehensive snapshot of the multifaceted emotions expressed by Twitter users.

Foundations for Further Discussion:

These findings lay robust foundations for subsequent in-depth discussions and contextual comprehension of public perspectives surrounding this globally significant sporting event. The word cloud, acting as a visual synthesis of sentiments, serves as a launchpad for more detailed analyses and discussions surrounding the diverse viewpoints prevalent in the online discourse.

Looking Ahead:

As future research initiatives unfold, there exists an opportunity to delve into refining sentiment analysis models. Furthermore, exploring additional contextual factors could further enrich the accuracy and depth of insights derived from the dynamic landscape of social media discourse. This forward-looking approach aims to continually enhance our understanding of the evolving sentiments and opinions surrounding major global events like the 2023 Qatar World Cup.

CONCLUSION

This research delves into the sentiments of Twitter users regarding the 2023 Qatar World Cup using sentiment analysis methods such as TF-IDF weighting and classification algorithms like Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), and Random Forest (RF) on Google Colab. The model evaluation results indicate that each algorithm provides unique insights into its ability to classify sentiments, although specific accuracy and

*name of corresponding author

F1 score figures are not detailed. Despite the study's limitations, focusing on Twitter data with Qatar 2023 World Cup keywords and the selection of three sentiment categories, the findings offer a profound understanding of public perspectives. The interpretive test's word cloud visualizations provide a rich context regarding dominant sentiments among Twitter users. Overall, this research significantly contributes to understanding public sentiments toward the Qatar 2023 World Cup. Model evaluation offers relative insights into algorithm performance, while word clouds provide an effective visual approach to summarizing prominent opinions and sentiments. These findings establish a strong foundation for further discussions and contextual understanding of public perceptions of this global sporting event, providing impetus for future research in sentiment analysis.

REFERENCES

Asro'i, A., & Februariyanti, H. (2022). Analisis Sentimen Pengguna Twitter Terhadap Perpanjangan Ppkm Menggunakan Metode K-Nearest Neighbor. *Jurnal Khatulistiwa Informatika*, *10*(1), 17–24. https://doi.org/10.31294/jki.v10i1.12624

Dedi Darwis, Nery Siskawati, & Zaenal Abidin. (2020). Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional. *Jurnal TEKNO KOMPAK*, *15*(1), 131–145.

Dewi, S., & Arianto, D. B. (2023). Twitter Sentiment Analysis Towards Qatar as Host of the 2022 World Cup Using Textblob. *Journal of Social Research*, *2*(2), 443–455. https://doi.org/10.55324/josr.v2i2.615

Dharmawan, L. R., Arwani, I., & Ratnawati, D. E. (2020). Analisis Sentimen pada Sosial Media Twitter Terhadap Layanan Sistem Informasi Akademik Mahasiswa Universitas Brawijaya dengan Metode K- Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(3), 959–965. http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/7099

Furqan, M., Sriani, S., & Sari, S. M. (2022). Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia. *Techno.Com*, *21*(1), 51–60. https://doi.org/10.33633/tc.v21i1.5446

Homepage, J., Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). IJCIT (Indonesian Journal on Computer and Information Technology) Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2), 118–127.

Isnain, A. R., Supriyanto, J., & Kharisma, M. P. (2021). Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(2), 121. https://doi.org/10.22146/ijccs.65176

Kristiyanti, D. A., Normah, & Umam, A. H. (2019). Prediction of Indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis. *Proceedings of 2019 5th International Conference on New Media Studies, CONMEDIA 2019*, 36–42. https://doi.org/10.1109/CONMEDIA46929.2019.8981823

Maulana, R., Voutama, A., & Ridwan, T. (2023). Analisis Sentimen Ulasan Aplikasi MyPertamina pada Google Play Store menggunakan Algoritma NBC. *Jurnal Teknologi Terpadu*, *9*(1), 42–48. https://doi.org/10.54914/jtt.v9i1.609

Normah, Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. *Jurnal Teknik Komputer AMIK BSI*, *8*(2), 174–180. https://doi.org/10.31294/jtk.v4i2

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 697–711. https://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369/348

Novantika, A. (2022). Analisis Sentimen Ulasan Pengguna Aplikasi Video Conference Google Meet menggunakan Metode SVM dan Logistic Regression. *PRISMA, Prosiding Seminar Nasional Matematika*, *5*, 808–813. https://journal.unnes.ac.id/sju/index.php/prisma/

Nurhakim, N., Handayanna, F., & Rinawati, R. (2017). Sistem Pakar Diagnosa Autisme Pada Anak Berbasis Android. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *1*(2), 158. https://doi.org/10.30645/j-sakti.v1i2.38

Ratna, S. (2020). Pengolahan Citra Digital Dan Histogram Dengan Phyton Dan Text Editor Phycharm. *Technologia: Jurnal Ilmiah*, *11*(3), 181. https://doi.org/10.31602/tji.v11i3.3294

Ray, S., Alshouiliy, K., & Agrawal, D. P. (2021). Dimensionality reduction for human activity recognition using google colab. *Information (Switzerland)*, *12*(1), 1–23. https://doi.org/10.3390/info12010006

Rijali, A. (2018). *Analisis Data Kualitatif Ahmad Rijali UIN Antasari Banjarmasin*. *17*(33), 81–95.

Wati, R., Ernawati, S., & Rachmi, H. (2023). Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH. *Jurnal Manajemen Informatika (JAMIKA)*, *13*(1), 84–93. https://doi.org/10.34010/jamika.v13i1.9424

*name of corresponding author