Journal of Social Research

P-ISSN: 2827-9832
E-ISSN: 2828-335x
http://ijsr.internationaljournallabs.com/index.php/ijsr

TWITTER SENTIMENT ANALYSIS TOWARDS QATAR AS HOST OF THE 2022 WORLD CUP USING TEXTBLOB

Syarafina Dewi, Dede Brahma Arianto
Satya Wacana Christian University, Islamic University of Indonesia
672019063@student.uksw.edu, 18917109@students.uii.ac.id

ABSTRACT

Twitter provides services to its users in the form of creating status messages or what is usually called tweets. Through tweets, users can express opinions, views, or emotions toward a topic. On 2 December 2010, Qatar was selected to host the 2022 World Cup. Qatar's selection as the host of the 2022 World Cup could elicit a variety of responses from various circles around the world. This research used TextBlob to find out the sentiment of Twitter users around the world regarding Qatar as the host of the 2022 World Cup. The research uses three stages, the first stage before Qatar was selected to host there was 88.46% positive sentiment and 11.54% negative sentiment, the second stage after Qatar was selected to host there was 79.38% positive sentiment and 20.62% negative sentiment, the third sentiment when the 2022 World Cup took place in Qatar was 83.72% positive sentiment and 16.28% negative sentiment. Based on the study, an accuracy score of 83% was obtained, meaning that the model was able to accurately predict 83% of the total testing data. This study can predict new data without having to be labeled first.

Keywords: *Qatar, FIFA World Cup 2022, sentiment analysis, TextBlob*

INTRODUCTION

Twitter is one of the most popular social communication platforms today. Twitter can connect people around the world through computers or mobile phones. Twitter provides services to its users in the form of creating status messages or what is usually called tweets (Fikri et al., 2020). Through tweets, users can express opinions, views, or emotions towards a topic, if the topic gets a lot of responses accompanied by the use of hashtags by users it will become a trending topic.

One of the topics that echo on Twitter and is a trending topic right now is the 2022 FIFA World Cup. On 2 December 2010, Qatar surprisingly won the bid as host of the 2022 FIFA World Cup, thus Qatar becoming the first Middle Eastern country to host the World Cup. Qatar was chosen to host after defeating Australia, Japan, South Korea, and the United States who are also running to host the 2022 World Cup. Qatar's selection as the host of the 2022 World Cup could elicit a variety of responses from various circles around the world. Social media such as Twitter is one of the places to give a response or opinion. (Odd, 2016) This can be used as material for sentiment analysis toward Qatar as the host of the 2022 World Cup.

Sentiment analysis belongs to one of the areas of Natural Language Processing (NLP) and is a process designed to identify the content of datasets in the form of opinions or views (sentiments) in the form of text on topics that are positive, negative, or neutral. However, users will face difficulties when reading Fauzi & Adinugroho (2018) tweets directly without marking them as positive, negative, or neutral. Therefore, a classification is needed that allows users to easily see which tweets are of positive or negative value (Primary, Ariesta, & Gata, 2022).

Research conducted by Ravikumar Patel and Kalpdrum Passi entitled "Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning" explains that the analysis was carried out by applying machine learning techniques. The data collected in this study is tweet data with hashtags "#brazil2014", "#worldcup2014", and match hastags. Results showed that Naïve Bayes provided the best accuracy of 88.17%, while random forest provided the best area with an AUC of 0.9 (Patel & Passi, 2020).

Research conducted by I Gede Susrama Mas Diyasa et al with the title "Twitter Sentiment Analysis as an Evaluation and Service Base on Python Textblob" explained that Twitter can be a place to express feelings, to customers, one of which is PT Telkom Indonesia. Tweet data processed using Textblob resulted in 34.4% positive tweets, 16.1% negative tweets, and 49.6% neutral tweets (Mas Diyasa et al., 2021).

METHOD

This study used the flowchart methodology in Figure 1.

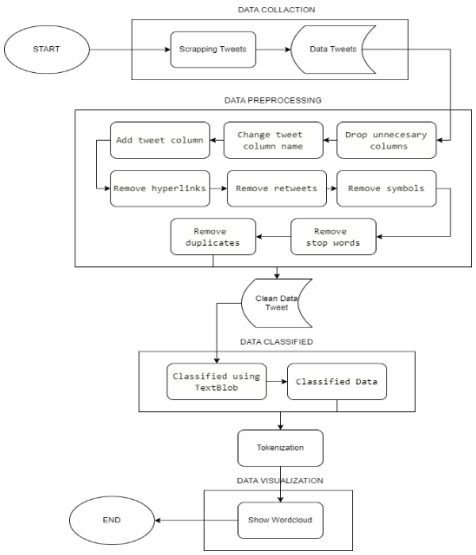

Figure 1. Research methodology flowchart

The data that has been collected cannot be directly analyzed because the data is still not clean. Data that has not been cleaned must go through several stages of Data Preprocessing. Once the data is clean, it will be analyzed using TextBlob to find out which tweets are positive and negative. The final step is the visualization of the results of the analysis.

1. Data Collection

Tweet data is captured by means of web scraping using SNScrape. SNScrape is a python library that is useful for web scraping on social media, one of which is Twitter. Researchers took a Rivaldi et al (2022) tweet in English relating to the World Cup in Qatar in 2022. Data is collected starting before Qatar was selected to host, after being selected to host, and when the 2022 World Cup took place in Qatar.

2. Data Preprocessing

Data Preparation is performed before the dataset is implemented into the model. Data Preparation is a must. The goal is to reduce data noise to make it cleaner so that maximum results are obtained. The process is divided into several stages, namely: (1) drop unnecessary columns to remove unnecessary columns, (2) change column name to change column names

to make it easier to understand, (3) add tweet columns to add new columns that later the data has been cleaned, (4) remove hyperlinks to remove links in tweets, (5) remove retweets to remove the retweet label in text, (6) remove symbols to remove symbols in the tweet, (7) remove stopwords to remove words that have a function but have no meaning, (8) remove duplicate to delete duplicate data (Giovani et al., 2020).

3. Data Classified

The sentiment analysis process is performed using the TextBlob library. TextBlob is a library used for textual information that provides a simple API for accessing Neuro-Linguistic Programming (NLP) activities. (Hazarika et al., 2020)The TextBlob library is capable of processing three types of classifications, namely positive, negative, and neutral. However, TextBlob can only be done in English. Therefore, researchers only take tweets that use English. In TextBlob there is a calculation that returns the polarity value. If the polarity value is greater than 0, then it is positive. If the polarity value is equal to 0, it is neutral. If the polarity value is less than 0, then negative (Mas Diyasa et al., 2021).

4. Tokenization

Tokenization is the process of breaking a sentence into a stand-alone set of words. Tokenization breaks down text that was originally a sentence into words. Tonization also eliminates delimiters such as periods, commas, spaces, and number characters in sentences (Hafidz, 2020).

5. Data Visualization

The visualization of the sentiment analysis results is displayed with the matplotlib library which is one of the libraries in Python for performing statistical data processing, visualization, and plotting (LEMENKOVA, 2019). The visualization is displayed in the form of a word cloud. A Word cloud is a display of words that appear frequently, and its size shows the frequency of occurrence in the data. Words whose frequency of occurrence is larger in size. On the contrary, words whose frequency of occurrence is less are smaller in size (Permadi, 2020).

RESULTS AND DISCUSSION

1. Data Collection

The web scraping process is carried out in three stages, the first stage was when Qatar offered to host in March 2009 until before the announcement of Qatar's selection to host on December 1, 2010, the second stage which was after the official announcement of Qatar's selection to host on December 2, 2010, until one year later on December 2, 2011, and the last stage is when the FIFA World Cup starts on November 20, 2022 until December 18, 2022.

Table 1. Twitter Data Descriptions

Keywords	Date	Sum
Qatar 2022 World Cup	March 1, 2009 - December 1, 2010	3606
Qatar 2022 World Cup	December 2, 2010 - November 19, 2022	41225
Qatar 2022 World Cup	November 20, 2022 - December 18, 2022	200001

Table 1 is the date the tweet was retrieved with the amount of data on that date. The maximum data results from web scraping are up to 200,000 at each stage, this is done so that the data is not unlimited.

Table 2. Sample Tweet Data

Datetime	Tweet Id	Text	Username
2010-11-30	9.72454E+15	Handicapping the 2022 World Cup Candidates - Qatar – First of all, Qatar is HOT in the summer, like kill... http://tumblr.com/xfzxxnsrs	jeff_underscore
2010-11-30	9.72392E+15	Aussies 2022 World Cup bid takes a hit report ranks Aus last in revenue USA scored 100%, AUS 68%, Japan 73%, Sth Korea 71%, Qatar 70%	meredith_gibbs
2010-11-30	9.72213E+15	Branding The #Qatar World Cup 2022 Bid MT @RobaAssi: http://bit.ly/i9QfTS via @natashaTynes	octavianasr
2010-11-30	9.71964E+15	Qatar 2022 World Cup Bid Gets Boost from Facebook – Mediabistro ... http://bit.ly/fOn8VA #worldcup	DC_Cappers
2010-11-30	9.71433E+15	Spector scored those two goals in order to get back at Sir Alex Ferguson backing Qatar's 2022 World Cup Bid. #goUSAbid	ChrisThomasFC

In this study, researchers used the attributes 'Datetime', 'Tweet Id', 'Text', and 'Username' taken from the web scraping process. An example sample of web scraping result data is shown in Table 2. After the data is collected, the tweet frequency can be seen in Figure 2, Figure 3, and Figure 4.

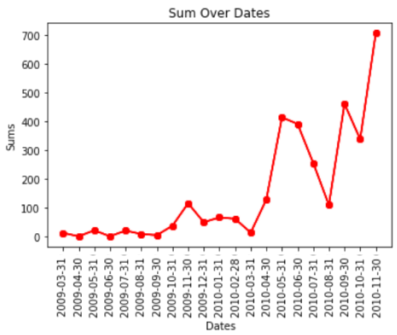

Figure 2. Tweet Frequency Before Qatar's Vote to Host

The diagram in Figure 2 shows the frequency of tweets from the time Qatar was nominated to host the 2022 FIFA World Cup in March 2009 until before the announcement Qatar was officially selected to host the 2022 FIFA World Cup on December 1, 2010. It can be seen that after March 2010 there was a rapid increase in the number of tweets although there was a decrease, but after that the number of tweets rose again.

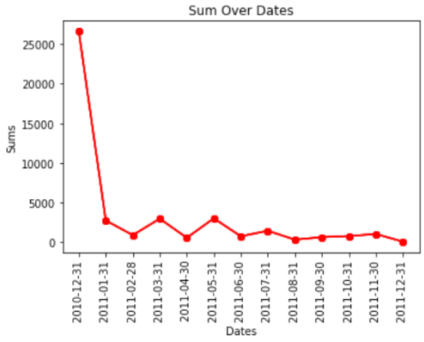

Figure 3. Tweet Frequency After Qatar Was Selected to Host

The diagram in Figure 3 shows the frequency of tweets after Qatar was selected to host the 2022 FIFA World Cup on December 2, 2010, to one year later on December 2, 2011. It can be seen that the number of tweets is the most right after the announcement, there is a sharp decline in January 2011, then the chart tends to stabilize.

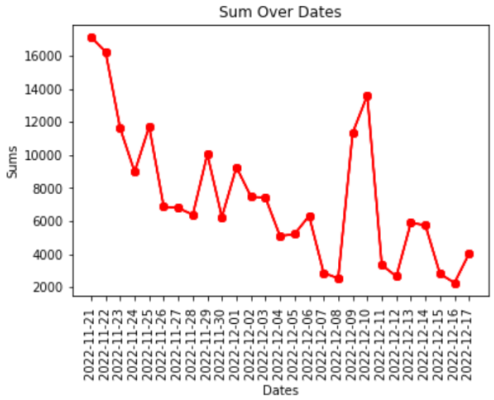

Figure 4. Tweet Frequency during the 2022 FIFA World Cup

The diagram in Figure 4 shows the frequency of tweets at the start of the 2022 FIFA World Cup in Qatar from November 20, 2022, to December 18, 2022. It can be seen that the peak of tweets occurred at the opening of the World Cup, after that there was a decline. Even though the number of tweets had increased quite drastically on December 9 to 10 2022, after that it went down again.

2. Data Preprocessing

Table 2 contains sample data, the 'Original Tweet' column shows tweets that still contain unnecessary URLs, usernames, and ASCII symbols that could interfere with the visualization. Therefore, it must be removed first. The data that has not been cleaned is carried out in several processes that have been described in the research methods section. After several stages of data preprocessing, the data will be easier to process into information. The clean data can be seen in Table 3 of the 'Tweets' column.

Table 3. Sample Data Before Cleaning and After Cleaning

Original Tweet	Tweets
Aussies 2022 World Cup bid takes a hit report ranks Aus last in revenue USA scored 100%, AUS 68%, Japan 73%, Sth Korea 71%, Qatar 70%	Aussies World Cup Bid takes Hit Report Ranks as last revenue USA scored as Japan ST Korea Qatar
Branding The #Qatar World Cup 2022 Bid MT @RobaAssi: http://bit.ly/i9QfTS via @natashaTynes	branding Qatar world cup bid mt via
Qatar bidding nation FIFA world cup 2022	Qatar bidding nation FIFA world cup
Let's all support #qatar and its world cup 2022 bid http://yfrog.com/47jlp01j	Let Us Support Qatar World Cup Bid
Is it really smart for USA fans to make fun of the Qatar 2022 World Cup bid based on their human rights record? #JustSaying	really smart USA fans make fun Qatar world cup bid based human rights record just saying

3. Data Classified

The data structure that used in this process is data that already through the data preprocessing process. Tweet text data consist of predictor variable that is tweets that already cleaned, and response variable contains the tweet sentiment of the classification results (positive and negative). Next, code to implementation TextBlob.

```
#Use textblob's sentiment method to analyze sentiment of passed tweet
after_tweets_df[['Polarity',  'Subjectivity']] = after_tweets_df['Tweet'].apply(Lambda Text: pd. Series(TextBlob(Text).sentiment))

#Calculating Negative, Positive and Compound values
for index, row in after_tweets_df'Tweet'].iteritems():
score = SentimentIntensityAnalyzer().polarity_scores(row)
neg = score['neg']
post = score['post']
comp = score['compound']

if neg > post:
after_tweets_df.loc[index,'Sentiment'] =  "-1"
else:
after_tweets_df.loc[index,'Sentiment'] =  "1"

after_tweets_df.loc[index,'neg'] = neg
after_tweets_df.loc[index,'post'] = post
after_tweets_df.loc[index,'compound'] = comp
```

The sentiment result of ‘1’ is positive, while the result of sentiment of ‘-1’ is negative. If the negative score exceeds the positive score, then the tweet is included in negative sentiment, besides that it is included in positive sentiment. This can be seen in the code. Table 4 is a sample of the results of the analytical sentiment used in this study.

Table 4. Sample Sentiment Analysis Results

Original Tweets	Tweets	Polarity	Subjectivity	Sentiment	Neg	post	compound
Let's all support #qatar and its world cup 2022 bid http://yfrog.com/47jlp01j	Let us support qatar world cup bid	0	0	1	0	0.31	0.4
@kyle_newman I assume de Qatar sheiks will bribe more than the US for the 2022 World Cup. Besides in that region there was no WC before.	Assume de Qatar Sheiks Bribe US World Cup besides Region WC	0	0	-1	0.15	0	-0.2
Excited for the world cup bid results, I wish, from the bottom of my heart that Qatar will host it in 2022 #Qatar2022	Excited World Cup Bid Results Wish Bottom Heart Qatar Host Qatar	0.375	0.75	1	0	0.36	0.62
FIFA Qatar Eyes 2022 World Cup http://tinyurl.com/3x7tbna	fifa qatar eyes world cup	0	0	1	0	0	0
Mark it: Russia for World Cup 2018, Qatar for World Cup 2022. Nothing surprises me with FIFA anymore. Lack of FIFA talk for US bid is odd.	mark russia world cup qatar world cup nothing surprises fifa anymore lack fifa talk us bid odd	-0.1667	0.25	-1	0.31	0	-0.64

	Total	Percentage
1	2277	88.46
-1	297	11.54

Figure 5. Percentage of Sentiment Before Qatar's Election to Host

The classified data process is divided into three stages, the first stage being when Qatar offered to host in March 2009 until before the announcement of Qatar's selection to host on December 1, 2010. Figure 5 is the result of the sentiment analysis produced in the first stage. Before Qatar was announced as the host, it showed that people had little negative feelings towards Qatar. In contrast, positive sentiment resulted is more than 88% of the total data retrieved. The number of tweets before Qatar was announced as the host was also very small compared to after the announcement and at the time Qatar's 2022 World Cup took place.

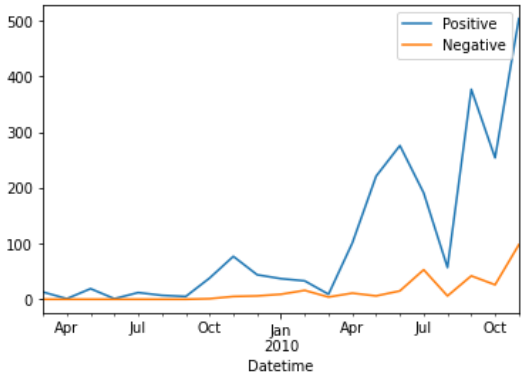

Figure 6. Comparison of Positive and Negative Sentiment before the Announcement

In Figure 6 the Y axis represents the number of tweets. Meanwhile, the X axis represents each month. The line chart compares negative sentiment and positive sentiment each month.

	Total	Percentage
1	23147	79.38
-1	6013	20.62

Figure 7. Percentage of Sentiment After Qatar's Election to Host

The second stage was after the official announcement of Qatar's election to host on December 2, 2010 until one year later on December 2, 2011. Figure 7 is the result of the sentiment analysis produced in the second stage. After Qatar was announced as the host, it shows that people have an increase in negative sentiment toward Qatar. Meanwhile, positive sentiment has decreased slightly from before. When viewed from the number of tweets after Qatar was announced as the host has increased a lot compared to before Qatar was announced as the host. The increase in the number of tweets was followed by an increase in negative sentiment towards Qatar.

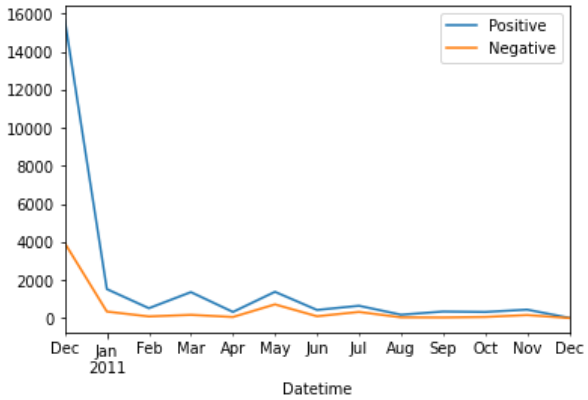

Figure 8. Comparison of Positive and Negative Sentiment after the Announcement

In Figure 8 the Y axis represents the number of tweets. Meanwhile, the X axis represents the number of tweets in each month. The line chart compares negative sentiment and positive

sentiment each month. The increase in positive sentiment was accompanied by an increase in negative sentiment.

	Total	Percentage
1	139147	83.72
-1	27051	16.28

Figure 9. Sentiment Percentages during the 2022 FIFA World Cup

The third stage is when the 2022 FIFA World Cup in Qatar starts on November 20, 2022, until December 18, 2022. Figure 9 is the result of the sentiment analysis produced in the third stage. During the 2022 FIFA World Cup in Qatar, people have decreased negative sentiment towards the 2022 World Cup in Qatar. Otherwise, positive sentiment has increased than after Qatar was announced as the host. When viewed from the number of tweets during the 2022 FIFA World Cup in Qatar has increased dramatically than after Qatar was announced as the host.

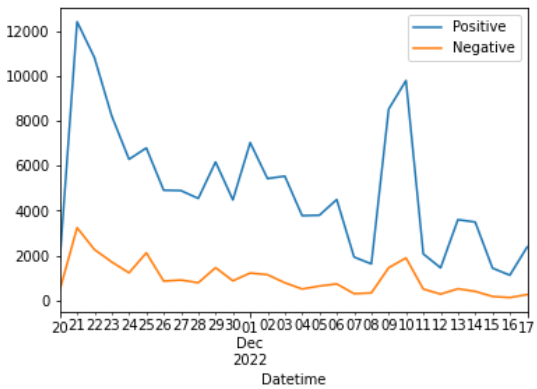

Figure 10. Comparison of Positive and Negative Sentiment during the 2022 World Cup

In Figure 10 the Y axis represents the number of tweets. Meanwhile, the X axis represents each day. The line chart compares negative sentiment and positive sentiment each day. The increase in positive sentiment was accompanied by an increase in negative sentiment.

4. Tokenization

The tokenization process is used to recognize words and break sentences into words based on spaces and punctuation. Table 4 is a sample of the tokenization results.

Table 5. Tokenization Process

Tweets	Tweets
branding qatar world cup bid mt via	'branding', 'qatar', 'world', 'cup', 'bid', 'mt', 'via'
qatar bidding nation fifa world cup	'qatar', 'bidding', 'nation', 'fifa', 'world', 'cup'
Let Us Support Qatar World Cup Bid	'let', 'us', 'support', 'qatar', 'world', 'cup', 'bid'

5. Data Visualization

The data that has gone through the tokenization process is then grouped to get the frequency of the words that appear the most. The frequency of this word will be visualized into a positive and negative wordcloud as shown in Figure 11, Figure 12, and Figure 13.

Figure 11. Positive and Negative Wordcloud Before Qatar Was Selected to Host

Figure 12. Positive and Negative Wordcloud After Qatar Was Selected as Host

Figure 13. Wordcloud Positives and Negatives when the 2022 World Cup takes place in Qatar

6. Confusion Matrix Calculation

Confusion matrix is a method used to measure the performance of a model. The Confusion Matrix contains information that compares the results of the classification made by the model with the actual results. There are four terms to represent the results of the classification process, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). A True Positive value is the amount of positive data that is correctly classified by the model.

A True Negative value is the amount of negative data that the model correctly classifies. The value of False Positive is the sum of positive data that are misclassified by the model. A False Negative value is the amount of negative data that are misclassified by the model (Setiawan et al., 2020).

```
Accuracy Score: 83.35%
---------------------------------------------------------
Confusion Matrix

    predicted_positive  predicted_negative
 1               22164                   0
-1                4428                   0
---------------------------------------------------------
Classification Report

              precision    recall  f1-score   support

          -1       0.00      0.00      0.00      4428
           1       0.83      1.00      0.91     22164

    accuracy                           0.83     26592
   macro avg       0.42      0.50      0.45     26592
weighted avg       0.69      0.83      0.76     26592
```

Figure 14. *Confusion Matrix* Results

This research performs calculations that include accuracy and precision. The accuracy value indicates how accurately the system can correctly classify the data. The precision value indicates the amount of data correctly classified as positive, divided by the total positive data. The results of the confusion matrix are shown in Figure 14. The test result of 80% of the training data and 20% of the overall testing data. TextBlob sentiment analysis yielded an accuracy of 83.35%.

CONCLUSION

In this study, sentiment analysis of tweets towards the 2022 FIFA World Cup in Qatar has been carried out using TextBlob. Data obtained from Twitter through web scraping with the keyword search for the 2022 FIFA World Cup in Qatar, obtained data as many as 244,832 tweets. The data is then analyzed to determine positive and negative sentiment. The research uses three stages, the first stage before Qatar was selected to host there was 88.46% positive sentiment and 11.54% negative sentiment, the second stage after Qatar was selected to host there was 79.38% positive sentiment and 20.62% negative sentiment, the third sentiment when the 2022 World Cup took place in Qatar was 83.72% positive sentiment and 16.28% negative sentiment. Based on the study, an accuracy score of 83% was obtained, meaning that the model was able to accurately predict 83% of the total testing data. This study can predict new data without having to be labeled first.

REFERENCES

Fauzi, M. A., & Adinugroho, S. (2018). *Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking Tomato Ripeness Identification View project Text clustering View project*. https://www.researchgate.net/publication/322959527

Fikri, M. I., Sabrila, T. S., Azhar, Y., & Malang, U. M. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *Smatika Jurnal*, *10*(02), 71–76.

Ganji, S. K. (2016). Leveraging the World Cup: Mega Sporting Events, Human Rights Risk, and Worker Welfare Reform in Qatar. In *JMHS* (Vol. 4).

Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI. *Jurnal Teknoinfo*, *14*(2), 115. https://doi.org/10.33365/jti.v14i2.679

Hafidz, N., Anggraeni, S., Gata, W., Ilmu Komputer STMIK Nusa mandiri Jakarta, M., & Komputer STMIK Nusa Mandiri Jakarta, T. (2020). Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes. *JurnalJUPITER*, *12*(2), 1–11.

Hazarika, D., Konwar, G., Deb, S., & Bora, D. J. (2020). Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing. *Proceedings of the International Conference on Research in Management & Technovation 2020*, *24*, 63–67. https://doi.org/10.15439/2020km20

LEMENKOVA, P. (2019). Generic Mapping Tools and Matplotlib Package of Python for Geospatial Data Analysis in Marine Geology. *International Journal of Environment and Geoinformatics*, *6*(3), 225–237. https://doi.org/10.30897/ijegeo.567343

Mas Diyasa, I. G. S., Marini Mandenni, N. M. I., Fachrurrozi, M. I., Pradika, S. I., Nur Manab, K. R., & Sasmita, N. R. (2021). Twitter Sentiment Analysis as an Evaluation and Service Base On Python Textblob. *IOP Conference Series: Materials Science and Engineering*, *1125*(1), 012034. https://doi.org/10.1088/1757-899x/1125/1/012034

Patel, R., & Passi, K. (2020). Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. *IoT*, *1*(2), 218–239. https://doi.org/10.3390/iot1020014

Permadi, V. A. (2020). Analisis Sentimen Menggunakan Algritma Naïve Bayes Terhadap Review Restoran di Singapura. *Jurnal Buana Informatika*, *11*(2), 141–151. https://www.kaggle.com/hj5992/restaurantreviews

Pratama, A. E., Ariesta, A., & Gata, G. (2022). Analisis Sentimen Masyarakat terhadap Tim Nasional Indonesia pada Piala AFF 2020 Menggunakan Algoritma K-Nearest Neighbors The researcher uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) method and implements the K-Nearest. *Jurnal TICOM: Technology of Information and Communication*, *10*(3), 187–196.

Rivaldi, A. A., Azra, B., Ziaulhaq, Y. I., & Rakhmawati, N. A. (2022). Analisis Karakteristik Akun Twitter Berdasarkan Sentimen Pendapat Terkait Undang-Undang PSE. *SATIN – Sains Dan Teknologi Informasi*, *8*(2), 25–35. https://doi.org/10.33372/stn.v8i2.876

Setiawan, A., Diyasa, I. G. S. M., Hatta, M., & Puspaningrum, E. Y. (2020). Mixture gaussian v2 based microscopic movement detection of human spermatozoa. *International Journal of Advances in Intelligent Informatics*, *6*(2), 210–222. https://doi.org/10.26555/ijain.v6i2.507