# SENTIMENT ANALYSIS ON THE TWITTER PSSI PERFORMANCE USING TEXT MINING WITH THE NAÏVE BAYES ALGORITHM

**Fajrullah Maulana[1]; M Arief Abdullah[2]; Juwita Sari[3]; Dimas Zappar Siddik[4]; Matius Agustinus[5]; Dedi Dwi Saputra[6]**

[1,2,3,4,5,6] Information Systems Study Program, Faculty of Technology and Information
Universitas Nusa Mandiri
Jakarta, Indonesia
www.nusamandiri.ac.id

[1]11220416@nusamandiri.ac.id, [2]11220475@nusamandiri.ac.id,
[3]11220438@nusamandiri.ac.id, [4]11220525@nusamandiri.ac.id, [5]11220521@nusamandiri.ac.id,
[6]dedi.eis@nusamandiri.ac.id
(*) Corresponding Author

**Abstract—**Social media has developed rapidly today, so social media is no longer just a place to interact and socialize but also to express opinions or criticize a particular party or institution. After the incident at the Malang Kanjuruhan stadium in October 2022, many netizens criticized the performance of PSSI as Indonesia's number one organization that oversees football competitions in Indonesia. For this reason, sentiment analysis was carried out on the official PSSI account on Twitter to assess the performance of PSSI by grouping them as Satisfied and Unsatisfied using the Naïve Bayes Classifier. Sentiment analysis took tweets from the official PSSI account and as many as 1000 comments to be used as a dataset. Then preprocessing is carried out in the GATA Framework using the Annotation Removal, Remove Hashtag, Transformation Remove URL, Regexp, Indonesian Steaming, and Indonesian Stopword Removal methods. The results obtained were 82.82% for accuracy, 78.69% for precision, 90.33% for recall, and 0.866 for AUC. With these results, the value obtained is at a good classification level.

**Keywords**: Sentiment Analysis, PSSI, Twitter, Naïve Bayes Classifier

***Abstrak—****Sosial media sudah berkembang sangat pesat di zaman sekarang ini sehingga sosial media tidak lagi hanya menjadi tempat untuk berinteraksi dan bersosialisasi namun juga sebagai tempat untuk menyampaikan pendapat ataupun kritik terhadap suatu pihak atau instansi tertentu. Pasca kejadian yang terjadi di stadion Kanjuruhan Malang pada Oktober 2022, banyak warganet yang mengkritisi kinerja dari PSSI sebagai Organisasi nomor satu Indonesia yang menaungi kompetisi sepakbola di Indonesia. Atas dasar itulah dilakukan sentimen analisis di akun resmi PSSI di Twitter untuk menilai kinerja dari PSSI dengan pengelompokkan Puas dan Tidak Puas dengan menggunakan algoritma Naïve Bayes. Analisis Sentimen dilakukan dengan mengambil tweet dari akun resmi PSSI sebanyak 1000 komentar untuk dijadikan dataset. Kemudian dilakukan preprocessing di GATA Framework dengan metode Annotation Removal, Remove Hashtag, Transformation Remove URL, Regexp, Indonesian Steaming dan Indonesian Stopword Removal. Hasil yang didapatkan adalah 82,82% untuk accuracy, 78,69% untuk precision, 90,33%, untuk recall dan 0,866 untuk AUC. Dengan hasil tersebut maka nilai yang didapat berada di level good clasification*

***Kata Kunci****: Sentiment Analysis, PSSI, Twitter, Naïve Bayes Classifier*

## INTRODUCTION

Football will never be crowded and alive without its supporters. Sometimes support for this sport leads to violence. In football, supporters are additional players for every football team, but the presence of supporters can also be a double-edged sword. Besides increasing team spirit, it can also hurt the team through hooliganism. Football supporters or fans are spirits for football clubs and even become an identity of the city itself(Prastyawan, 2018)

Every sporting event or match needs the support of sports institutions to reduce the potential for negative behavior. The PSSI organization is the only national football organization with authority to regulate, manage and organize all football activities or competitions in Indonesia(Zulhidayat, 2018).

Sentiment analysis is extracting text data to obtain information about positive, neutral, or negative sentiments. Internet users on social media provide sentiment analysis to provide an

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 2 September 2022 | DOI: 10.33480/pilar.v18i2.3938

assessment or personal opinion(Sari & Wibowo, 2019). In this study, the authors analyzed the responses of Twitter media netizens to PSSI's performance after the tragedy at the Kanjuruhan Malang stadium using the RapidMiner application and the Naïve Bayes algorithm.

Referring to research conducted by(Lasepa et al., 2021) for the Indonesian National Team with the title "Sentiment Analysis of Netizens' Perspectives on the Kanjuruhan Malang Tragedy on Twitter Using the Naïve Bayes Classifier," where the research also uses the naïve Bayes algorithm. From this test, the results obtained for precision were 77.19%, recall results were 78.50%, accuracy results were 65.78%, and AUC was 0.820.

In research journals conducted by(Suryani et al., 2019) with the title "Use of the Naïve Bayes Classifier Method in Indonesian Language Facebook Sentiment Analysis," the steps for defining datasets, preprocessing, feature selection, labeling, classification, and evaluation, where the datasets used are 20 datasets for the initial test, and as many as 479 datasets at the end the test. The results obtained are an accuracy rate of 5%, an error of 95% at the beginning of the test, and 87.1% and 12.9% at the end. It is because the test data is similar to the system's training data. The more training data used in the system will impact system performance in the classification process.

Whereas in a research journal with the title "Analysis of Sentiment Against Public Opinion About the Covid-19 Vaccine Using the Naïve Bayes Classifier Algorithm"(Yulita et al., 2021) where the preprocessing stages of the research carried out was cleansing, converting negation, converting emoticons, case folding, tokenization, filtering stopwords and stemming in Indonesian with 2278 datasets, the final result was a positive response of 60.3%, a negative response of 5.4% and the neutral response of 34.4%.

Meanwhile, in research journals(Hermawan et al., 2022) with the title "Optimizing Sentiment Analysis on Twitter Olshop Tokopedia Using Textmining With the Naïve Bayes & Adaboost Algorithm" with 1000 datasets obtained results of accuracy of 94.95%, precision of 90.86%, recall of 100% and AUC of 0.950.

All the research above uses the Naïve Bayes algorithm by classifying positive and negative segments on social media such as Facebook, Instagram, and Twitter with a fairly high degree of accuracy.

## MATERIALS AND METHODS

The author's application in this study is RapidMiner using the Naïve Bayes algorithm through the data mining process. RapidMiner is a data science software platform developed by the company of the same name, which provides a unified environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications, research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process, including data preparation, result visualization, validation, and optimization(Fadilah, 2019). Meanwhile, Naive Bayes is a simple probabilistic prediction technique based on the Bayes theorem with a strong assumption of independence(Imandasari et al., 2019). The advantage of Naive Bayes is that the data classification process can be adapted to the nature and needs of each (Gunawan et al., 2018). And data mining is an analysis of reviewing data sets to find unexpected relationships and summarize data in different ways that are understandable and useful to data owners(Utomo & Mesran, 2020)

In this study, the research method that the authors used is illustrated in Figure 1 below.

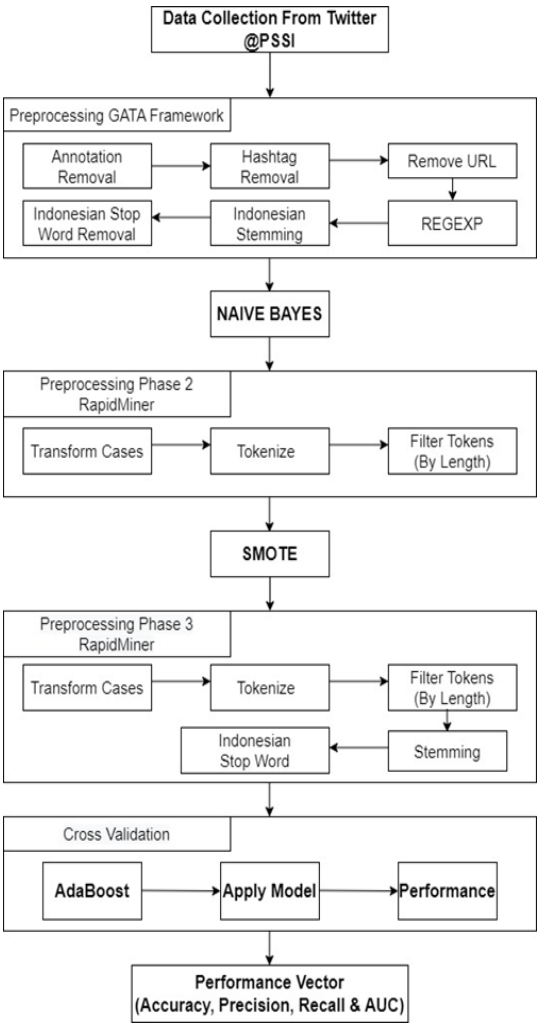



**Figure 1. Sentiment Analysis Framework**

The explanation of the framework above is as follows :

1. Data Collection From Twitter

This stage is the earliest stage of all existing processes. Researchers collect data from Twitter through RapidMiner by connecting RapidMiner to the account of a certain party or agency that has Twitter. As much data as possible is collected and stored in one excel file to be labeled by grouping satisfied and dissatisfied, positive and negative. Later this data will be used as a research dataset.

2. Preprocessing Phase 1

The data collected in the previous stage in excel form is entered into a machine learning framework to remove symbols, punctuation, or links so that the data becomes more structured data representing a dataset.

3. Preprocessing Phase 2

At this stage, the dataset processed in the machine learning framework will be processed again through the RapidMiner application. The operators used at this stage are Transform Cases, Tokenize, and Filter Token By Length, and they have also included the Naïve Bayes and SMOTE Upsampling operators. SMOTE or Synthetic Minority Oversampling Technique is needed to prevent imbalance because modeling with an algorithm that does not pay attention to data imbalance dominates the major class and does not pay attention to the minor class. Then run the RapidMiner application to see the results of the performance vector.

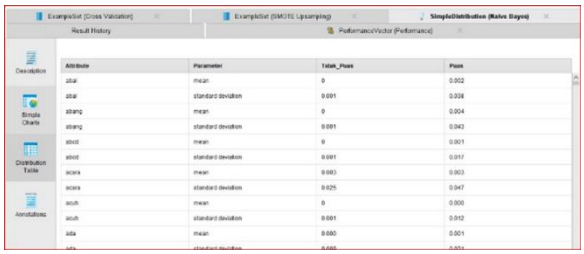


**Figure 2. Distribution Table**

Then in the distribution table like the figure 2 above, retrieve data in the Attribute column to prepare for the next stage.

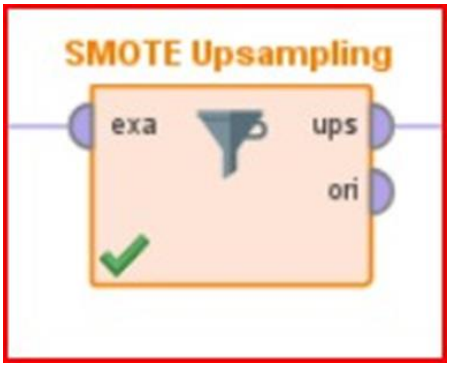



**Figure 3. SMOTE Operator**

SMOTE operator in figure 3 above is a derivative of over-sampling algorithm which is useful for increasing the sensitivity value to above 50%.

4. Preprocessing Phase 3

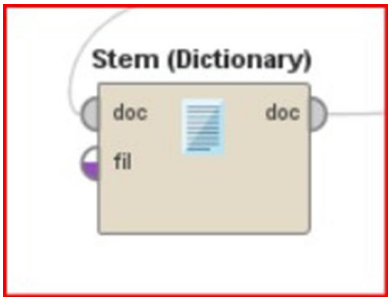



**Figure 4. Stem Operator (Dictionary)**

At this stage, an operator called Stem (Dictionary) like the figure 4 above is used, which functions to justify wrong words so that these words have a value in a text or sentence.

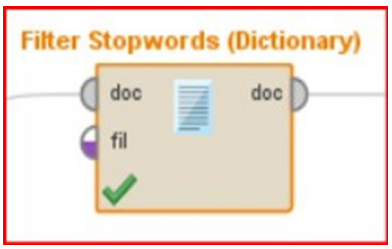



**Figure 5. Filter Stopwords Operator**

Then also at this stage, an operator called Filter Stopwords (Dictionary) like figure 5 above is also used where there is a notepad file containing data from the processed SMOTE Upsampling results, which removes words with low value or information from a text or sentence.

5. AdaBoost

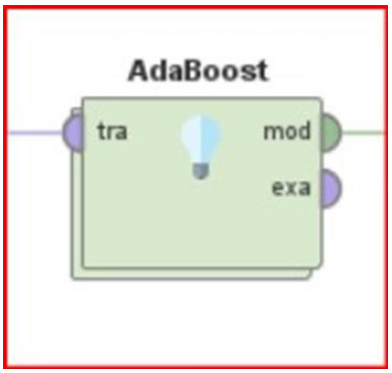



**Figure 6. AdaBoost Operator**

Adaboost operator like figure 5 above, is an ensemble learning that is often used in boosting algorithms. Boosting can be combined with other classifier algorithms to improve classification performance(Qadrini L et al., 2021). AdaBoost is

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 2 September 2022 | DOI: 10.33480/pilar.v18i2.3938

used to increase the confidence level of data for machine learning so that it can go to the next stage.

6. Confusion Matrix

**Table 1. Confusion Matrix**

| Correct Classic action | Classified as | |
|---|---|---|
| | + | - |
| + | True Positive | False Negatif |
| - | False Positif | True Negatif |

Confusion matrix in table 1 above is a table that states the classification of the number of correct test data and the number of erroneous test data(Normawati & Prayogi, 2021). The confusion matrix is usually used to calculate data mining concepts accurately (Mutawalli et al., 2019). The way the confusion matrix works is by comparing or combining 4 different values from the predicted value and the actual value.

The values in this matrix are accuracy, precision, recall, and AUC.

The accuracy value, actually the accuracy value, is the closeness level value between the predicted value and the actual value, Antinasari (in Syarifuddinn, 2020).

Precision value, namely the level of prediction accuracy of a system, by calculating positive predictions from the total data predicted by the system, including wrong predictions, Rahutomo(in Syarifuddinn, 2020).

The recall value is the success rate in recognizing a class that must be recognized, Hakiem & Fauzi(in Syarifuddinn, 2020).

AUC value (Area Under Curve), which is used to measure the difference in performance that has been calculated, Faisal (in Syarifuddinn, 2020).

## RESULTS AND DISCUSSION

1. Crawling Twitter Data

Collecting data from Twitter with the RapidMiner application. Connecting RapidMiner with the "Retrieve Connection-Twittter" operator. Then use the Twitter Search operator with the query @pssi, just like figure 7 below.

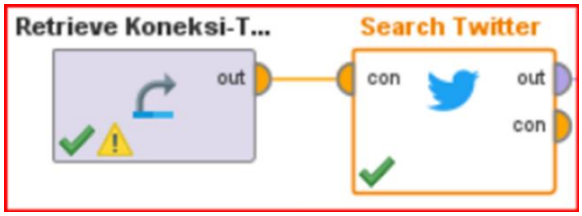
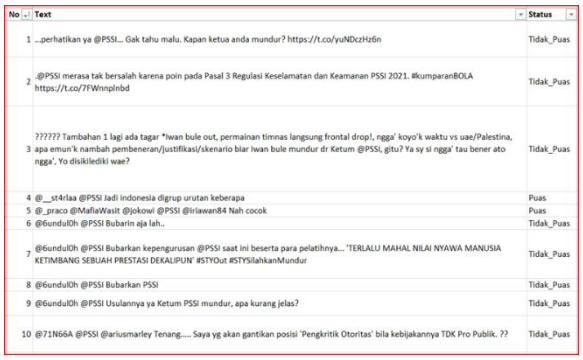



**Figure 7. Crawling Twitter Data**

There are 1456 data from the process above. The data from the crawling process is stored in a file in excel format. Then it is processed by eliminating twin or similar data so that the remaining 1000 data will be carried out in preprocessing phase 1.

2. Labelling

Labeling in figure 8 below is done by labeling each data with a predetermined class or category. The categorization of this data is "Satisfied and Dissatisfied."



| No | Text | Status |
|---|---|---|
| 1 | ...perhatikan ya @PSSI... Gak tahu malu. Kapan ketua anda mundur? https://t.co/yuNDczHz6n | Tidak_Puas |
| 2 | .@PSSI merasa tak bersalah karena poin pada Pasal 3 Regulasi Keselamatan dan Keamanan PSSI 2021. #kumparanBOLA https://t.co/7FWnnpInbd | Tidak_Puas |
| 3 | ?????? Tambahan 1 lagi ada tagar "Iwan bule out, permainan timnas langsung frontal drop!, ngga' koyo'k waktu vs uae/Palestina, apa emun'k nambah pembeneran/justifikasi/skenario biar Iwan bule mundur dr Ketum @PSSI, gitu? Ya sy si ngga' tau bener ato ngga', Yo disikilediki wae? | Tidak_Puas |
| 4 | @__st4rlaa @PSSI Jadi indonesia digrup urutan keberapa | Puas |
| 5 | @_praco @MafiaWasit @jokowi @PSSI @iriawan84 Nah cocok | Puas |
| 6 | @6undul0h @PSSI Bubarin aja lah.. | Tidak_Puas |
| 7 | @6undul0h @PSSI Bubarkan kepengurusan @PSSI saat ini beserta para pelatihnya... 'TERLALU MAHAL NILAI NYAWA MANUSIA KETIMBANG SEBUAH PRESTASI DEKALIPUN' #STYOut #STYSilahkanMundur | Tidak_Puas |
| 8 | @6undul0h @PSSI Bubarkan PSSI | Tidak_Puas |
| 9 | @6undul0h @PSSI Usulannya ya Ketum PSSI mundur, apa kurang jelas? | Tidak_Puas |
| 10 | @71N66A @PSSI @ariusmarley Tenang..... Saya yg akan gantikan posisi 'Pengkritik Otoritas' bila kebijakannya TDK Pro Publik. ?? | Tidak_Puas |

**Figure 8. Labelling**

Of the 1000 datasets, there are 287 data with satisfied labels and 713 with dissatisfied labels.

3. Preprocessing Phase 1

The labeled dataset is then processed in the GATA Framework using the @Annotation Removal technique, #(Hashtag) Removal Transformation Remove URL, Regexp, Indonesian Stemming, and Indonesian Stop Word Removal, just like figure 9 below.

**Figure 9. GATA Framework**

a. @Annotation Removal aims to remove the annotation mark on the tweet dataset because it is considered meaningless.
b. #(Hashtag) Removal, The goal is to delete words that start with a punctuation mark (#).
c. Transformation Remove URL, function to delete the URL address in the tweet dataset.
d. Regexp, remove symbols or emoticons in the tweet data set.
e. Indonesian Stemming aims to change the removal of affixes - words in the tweet dataset into basic words according to KBBI.
f. Indonesian Stop Word Removal aims to eliminate words not by KBBI and connect words such as "in" or "which." The stop-word process collects the words that appear most often in the corpus.

Initially, there were 1000 datasets, which decreased to 964 after preprocessing phase 1.

4. Preprocessing Phase 2

At this stage, the dataset is tested into the RapidMiner application using operators such as Naïve Bayes, SMOTE Upsampling, Transform Cases, Tokenize, dan Filter Tokens By (Length).

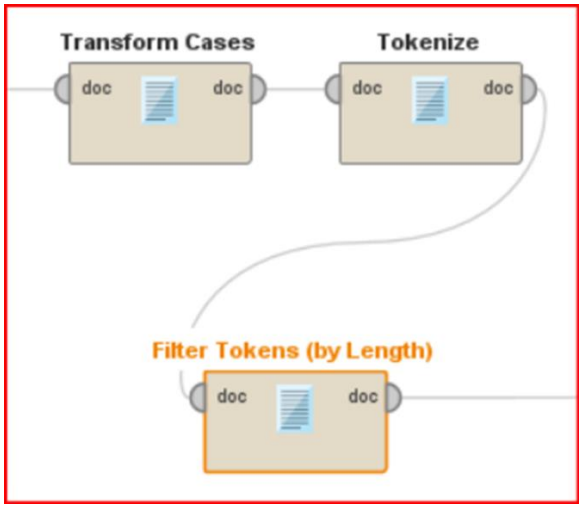



**Figure 10. Preprocessing Phase 2**

The explanation of the figure 10 above is as follows:

a. Transform Cases is used to convert all letters to lowercase or all uppercase.
b. Tokenize is the process of separating words. The process of slicing each word in the text and converting the letters in the document to lowercase. Only letters are accepted, and special characters or punctuation are omitted.
c. Filter Token (By Length) extracts important words from the resulting token. In this process, words that have a certain length will be deleted.

Words in the distribution table taken from RapidMiner, the results of preprocessing phase 2, are then collected for further processing.

5. Preprocessing Phase 3

At this stage, the words in the distribution table are taken, and then a stemming file is created in notepad by placing the correct word followed by the wrong word (correct word: wrong word). Then the stemming file is entered into the Stem operator (Dictionary).

After that, the words are weighted in the example dataset. Then the words that have been weighted are entered into a notepad file which later, the file is entered into an operator called Filter Stopword (Dictionary).

**Table 2. Stopword**

| Text | Value/Weight |
|---|---|
| jaya | 0,1589 |
| pusat | 0,1589 |
| sinergi | 0,1589 |
| zona | 0,1589 |
| hawa | 0,2054 |
| nafsu | 0,2054 |
| sementara | 0,2054 |

In table 2 above precisely in the Text column, it shows the words generated from the distribution table on rapidminer, while the column on the right shows the value of the word on the left, where the value is taken from Data on SMOTE Upsampling from rapidminer .The smaller the value or weight of a word, the more there is no correlation between the word and the research object.

6. AdaBoost

In the previous stages, an accuracy value of 82.12%, a precision of 77.68%, a recall of 90.33%, and an AUC of 0.709 were obtained. With these results, the AdaBoost operator is needed to increase the confidence of machine learning so that the AUC value can be above 0.80.

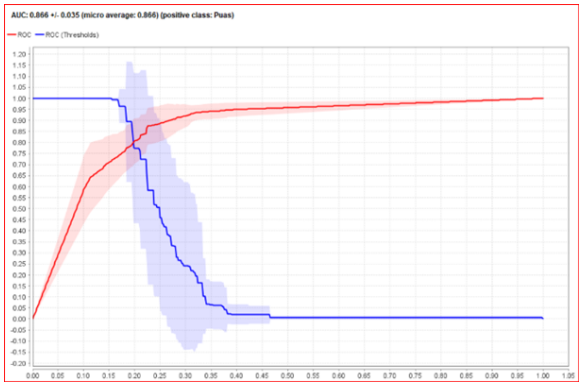


**Figure 11. ROC Graphic**

Figure 11 is the result of the last stage using AdaBoost, where the final results in this study experienced a significant increase. This increase is due to the use of AdaBoost itself, which functions to increase the confidence of machine learning data.

## CONCLUSION

Based on the research process described above, the results obtained were 82.82% for accuracy, 78.69% for precision, 90.33% & for recall, and 0.866 for AUC. Based on the criteria for the AUC rating range, the results obtained above are Included in the good classification and can continue to the deployment process.

## REFERENCE

Fadilah, E. (2019). Implementasi Metode Profile Matching Terhadap Sistem Pendukung Keputusan Penerimaan Dana Zakat pada Badan Amil Zakat Pertamina (BAZMA). Matics, 10(2), 39. https://doi.org/10.18860/mat.v10i2.5745

Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. Jurnal Edukasi Dan Penelitian Informatika (JEPIN), 4(2), 113. https://doi.org/10.26418/jp.v4i2.27526

Hermawan, D., Akhsanal, M., Wahyudi, Z., Ariyanto, A., & Dwi, D. (2022). Optimasi Analisis Sentimen Pada Twitter Olshop Tokopedia Menggunakan Textmining Dengan Algoritma Naïve Bayes & Adaboost. 6(September), 821–828.

Imandasari, T., Irawan, E., Windarto, A. P., & Wanto, A. (2019). Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air. Prosiding Seminar Nasional Riset Information Science (SENARIS), 1(September), 750. https://doi.org/10.30645/senaris.v1i0.81

Lasepa, R., Riyadi, S., Ramadhan, S., & Saputra, D. D. (2021). Analisis Sentimen Terhadap Perspektif Warganet Atas Tragedi Kanjuruhan Malang di Twitter Menggunakan Naïve Bayes Classifier. 8(1), 1–8.

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). KLASIFIKASI TEKS SOSIAL MEDIA TWITTER MENGGUNAKAN SUPPORT VECTOR MACHINE (Studi Kasus Penusukan Wiranto). Jurnal Informatika Dan Rekayasa Elektronik, 2(2), 43. https://doi.org/10.36595/jire.v2i2.117

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. J-SAKTI (Jurnal Sains Komputer Dan Informatika), 5(2), 697–711.

Prastyawan, E. B. (2018). Stereotip dan Konflik Antar Suporter Sepakbola Persibat dan Persip Pekalongan. Persepsi : Communication Journal, 1(1), 1–14. https://doi.org/10.30596/persepsi.v1i1.2440

Qadrini L, Sepperwali A, & Aina A. (2021). Decision Treedan Adaboostpada Klasifikasi Penerima Program Bantuan Sosial. Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial, 2(7), 1959–1966.

Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. Jurnal SIMETRIS, 10(2), 681–686.

Suryani, P. S. M., Linawati, L., & Saputra, K. O. (2019). Penggunaan Metode Naïve Bayes Classifier pada Analisis Sentimen Facebook Berbahasa Indonesia. Majalah Ilmiah Teknologi Elektro, 18(1), 145. https://doi.org/10.24843/mite.2019.v18i01.p22

Syarifuddinn, M. (2020). Analisis Sentimen Opini Publik Mengenai Covid-19 Pada Twitter Menggunakan Metode Naïve Bayes Dan Knn. INTI Nusa Mandiri, 15(1), 23–28. https://doi.org/10.33480/inti.v15i1.1347

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. Jurnal Media Informatika Budidarma, 4(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Yulita, W., Dwi Nugroho, E., Habib Algifari, M., Studi Teknik Informatika, P., Teknologi Sumatera, I., Terusan Ryacudu, J., Huwi, W., Agung, J., & Selatan, L. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. Jdmsi, 2(2), 1–9.

Zulhidayat, M. (2018). Kewenangan Dan Peran Pemerintah Dalam Penyelenggaraan Komepetisi Sepak Bola Di Indonesia (the Authority and Role of Government in the Organizing of Football Competition in Indonesia). Jurnal Hukum Replik, 6(2), 222. https://doi.org/10.31000/jhr.v6i2.1446