Journal of Applied Engineering and Technological Science
Vol 4(1) 2022 : 375-389

ANALYSIS OF TWITTER SENTIMENT TOWARDS MADRASAHS USING CLASSIFICATION METHODS

Supriadi Panggabean[1*] , Windu Gata[2], Tri Agus Setiawan[3]
Computer Science, Nusa Mandiri University Jakarta, Indonesia[12]
STIKOM Cipta Karya Informatika, Indonesia[3]
14002471@nusamandiri.ac.id

Received : 07 October 2022, Revised: 05 December 2022, Accepted : 05 December 2022
**Corresponding Author*

ABSTRACT
Several incidents of sexual violence, the emergence of radical Islamic issues, terrorism, intolerance of changes in the character of students and so on have recently become a highlight for madrasahs. To find out how the sentiment of social media users towards madrasahs, research on twitter sentiment towards madrasahs was conducted using text mining techniques. The methods used are Naïve Bayes (NB), Decision Tree (DT) and K – Nearest Neighbor (K-NN) which aim to classify public sentiment towards Madrasahs on Twitter. The dataset used is a tweet in Indonesian with the keyword "Madrasah" as many as 3288 tweets. The techniques used to build classification and sentiment analysis are text mining, transformation, tokenize, stemming and classification, etc. Gataframework tools, execute Python script and RapidMiner are also used to help create sentiment analysis in measuring classification values. The results obtained by the optimization using Particle Swam Optimization (PSO) using the Naïve Bayes algorithm and the accuracy value obtained was 80.80%, with a precision value of 83.03%, a recall value of 78.68%, and an AUC of 0.739.
Keywords : *Data Mining, Sentiment Analysis, Classification*

1. Introduction

In today's digital era, the influence and use of the internet has become a necessity, especially in Indonesian, internet users in Indonesia in early 2021 reached 202.6 million people. This number increased by 15.5 percent or 27 million people when compared to January 2020. The total population of Indonesia at this time is 274.9 million people. This means that internet penetration in Indonesia in early 2021 reached 73.7 percent. This is reported in a recent report released by content management service *HootSuite*, as well as social media marketing agency *We Are Social* in a report titled" Digital 2021". Internet activities that are very popular with Indonesian internet users are social media. Currently, there are 170 million Indonesians who are active users of social media. On average, they spend 3 hours and 14 minutes on the network platform social (Riyanto, 2021).

Social media that are often used in Indonesia include Instagram, Facebook, and Twitter. Although twitter is not as big as Facebook and Instagram at this time, it is sourced from data that the researcher read from tekno.kompas.com *published* on April 14, 2021 at 20.42 WIB reported that the development of Twitter at this time is getting better. In the first quarter of 2020, there was a surge in its daily active users from 134 million in the first quarter of 2019 to 166 million users or face an increase of 24 percent. In the second quarter of 2020, this figure increased again to 186 million users. The number of active users every day exceeded the forecasts of analysts, who initially estimated that they were only trying to reach 176 million users (Pratomo, 2021).

In Indonesia, Twitter users have unique characteristics compared to other countries. Indonesian users, use Twitter as a medium to express comments. Not only that, Twitter users also tend to notify the events that are intertwined around them. In a very short span of time an opinion or expression of a person will be so easy to see by many parties. Starting from that argument, there should be other arguments or opinions on the issue.

Many researchers also use social media as a reference for data taken for an important source of information about opinions, or community responses, and measure the level of popularity and become a benchmark for the services of these agencies, institutions, or companies. At this time the process of taking data through social media twitter where this platform is believed to be a platform whose opinions can and has value to be processed in several algorithms, it's just that to measure a comment sentiment on one of the social media is difficult.

Sentiment analysis can be interpreted as the process of extracting, processing, and understanding data automatically in the form of unstructured text to retrieve sentiment information contained in opinions or opinion sentences (Brahimi et al., 2019). The use of sentiment analysis to evaluate the trend of an opinion against negative opinions and positive opinions on a topic (Rozi et al., 2012). Sentiment analysis is a computational-based detection and learning of opinions or views (sentiments), emotions, and subjectivity in the text. As a special text mining application, sentiment analysis is related to the automatic extraction of positive or negative opinions from the text (He et al., 2015).

Text mining is one of the techniques that can be used to classify documents, where *text mining* is a variation of data mining that seeks to find interesting patterns from a large set of text data. One of the classification methods that can be used in doing *text mining* is the *Naive bayes* method. *Naïve Bayes* is a classification using probability and statistical methods (Suryanto et al., 2019) (Suryanto et al., 2019). The advantage of using Naïve Bayes is that this method only requires a small amount of training data to determine the estimated parameters required in the classification process. *Naïve Bayes* often works much better in most complex real-world situations than expected (S.A Pattekari, 2012). Naïve Bayes classifier's research intends to carry out the process of classifying the results of netizens' comments on the application of technology that has gone through a process of sentiment analysis. Another method used in this research is decision tree. Decision Tree is a very popular and practical approach in machine learning to solve classification problems(G. Wahyuningtyas, 2014). Apart from the fact that the construction is relatively fast, the results of the built model are easy to understand (Y. Sunoto, 2014). Next is the K-Nearest Neighbor (KNN) method which is often also used to analyze sentiment. The KNN method is the process of grouping data into predefined classes based on the closest distance/degree of similarity of that data to an existing dataset/training data(Deng & Yu, 2013).

Several incidents of sexual violence that occurred in the madrasa environment as reported in the media, the emergence of radical Islamic issues which he said were the fruit of thoughts from the madrasa environment, terrorism which was also said to come from misinterpreting knowledge from madrasahs, intolerance to different religions, changes in the character of madrasah students and so on will cause negative thoughts towards madrasahs. Until now, there has not been much research on sentiment towards madrasas. Based on the background that has been described, in this study raised the title of the thesis entitled " Analysis of Twitter Sentiment Towards Madrasahs Using Classification Methods". The method used to process data from twitter opinions, the author tries to use five methods as a comparison of which one is more accurate and can be processed data. These methods are Naïve Bayes (NB), Decision Tree (DT) and K – Nearest Neighbor (K-NN) using the RapidMiner application.

Problem Identification: Based on the description of the background of the problem above, in this study identifying problems that can be used as the object of research is how sentiment analysis of Twitter data regarding opinions on madrasas using methods are Naïve Bayes (NB), Decision Tree (DT) and K – Nearest Neighbor (K-NN). The purpose of writing this thesis is to get the best classifier in determining the classification of sentiment analysis on social media twitter Indonesian texts about Madrasah.

Scope of Research : In order for the discussion in this study to be more directed, the writing provides a limitation of the problem, namely: The sentiment category used includes positive sentiment and negative sentiment, the dataset used is data from social media twitter with Indonesian text that has a narrative related to Madrasah, The algorithm used for sentiment analysis is Naïve Bayes (NB), Decision Tree (DT) and K – Nearest Neighbor (K-NN) with a K-fold testing model Cross validation and compared results are only the results of the Accuracy pattern and AUC performance on the ROC curve to measure the model, to improve the performance of the classification method can be done using the Particle Swarm Optimization (PSO) feature selection and the data mining method used is the Cross Industry Standard Process for Data Mining (CRISP-DM).

2. Literature Review

Data Mining is a term used to describe the discovery of knowledge in a database. Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases(Turban, 2005). Data classification is a process that finds the same properties in a set of objects in a database and classifies them into different classes according to the established Classification model. Text Mining is mining carried out by a computer to obtain something new, something previously unknown or rediscover implicitly implied information, derived from information extracted automatically from different sources of text data Text mining is a technique used to deal with classification, clustering, information extraction and information retrieval problems (Xiaojun, 2011).

In data mining to measure or there are several ways to measure the performance of the resulting model, one of which is using a confusion matrix (accuracy). Confusion matrix is a method used to perform accuracy calculations on the concept of data mining. Precision or confidence is the proportion of positive predicted cases that are also positive in the actual data. Recall or sensitivity is the proportion of actual positive cases that are correctly predicted to be positive.

Table 1 - Model Confusion Matrix

Correct classification	*Classified as*	
	+	-
+	*True positive*	*False negative*
-	*False positive*	*True negative*

Source : (Ibrahim, 2017)

Sentiment analysis is extracting people's opinions, sentiments, evaluations, and emotions about a particular topic written using natural language processing techniques. A number of other major works mention sentiment analysis focusing on specific applications that classify positive, negative and neutral opinions(Alita et al., 2019). Sentiment analysis or also known as mining opinion is an analysis that aims to see the opinion of the community or group regarding certain entities (Safitri et al., 2021).

The preprocessing stage is needed to clean the data from unnecessary text, where the unstructured text data will be converted into structured or semi-structured text data. The stages of preprocessing to process data are case folding, convert emoticons, cleansing, tokenizing, stop word removal and stemming (Aditia Rakhmat Sentiaji et al., 2014)

Social media is a new set of communication and collaboration tools that enable many types of interactions that were previously unavailable to ordinary people. The most important thing about this technology is the shift in the way people know, read and share news, and search for information and content. There are hundreds of social media channels operating around the world today, with the top three on Facebook, LinkedIn, and Twitter (Dailey, 2009). Social media has several special characteristics including: Reach, Accessibility, Usability, Actuality and Permanently (Purnama, 2011). Twitter is the most popular microblogging in Indonesia. This microblogging allows users to send and read messages called tweets, in the form of a maximum of 140 characters of text displayed on the user's profile page(Badri, 2011).

The Naive Bayes approach is a classification method that refers to Bayes' theorem. Bayes' theorem is used to calculate the probability of data uncertainty (Peter Norvig, 2010). The Naïve Bayes Classifier approach process assumes that the presence or absence of a feature in a class is not related to the presence or absence of other features in the same class (Setiawan et al., 2021).

The equation of Bayes' theorem is $P(H|X) = \frac{P(H|X).P(H)}{P(X)}$ (Muktamar et al., 2015).

The Decision Tree is a tree-like flowchart structure, where each internal node represents an attribute test, each branch represents the test result, and the leaf node represents a class or class distribution(Kasih, 2019).

The K-Nearest Neighbor (KNN) method is the simplest of all other classification methods for solving classification problems. The technique used in this K-NN is to classify the data using objects with adjacent closest values. The results obtained from this process are higher or best when the weighting of *the similarity of Cosine Similarity* is used in the calculation of each tribe. Text classification with the K-NN method gives a better value when the *expression Cosine Similarity* is used to weigh each word in the text document being processed before calculating the value of *Cosine Similarity*, after the word weighting is completed the steps of the word weighting process are carried out, namely tf, df, idf, tfidf, and use the *Cosine Similarity* formula to perform similarities between documents (Nurjanah et al., 2017).

Particle Swarm Optimization (PSO) is often used in research, because PSO has similar properties to *genetic algorithms (GA)*. The advantage of PSO is that it is easy to implement and there are several parameters to adjust. The PSO system is initiated by a random solution population and then finds the optimum point by updating each generation result. The approach used is more systematically mathematical to find solutions. *Particle Swarm Optimization (PSO)* was formulated by Edward and Kennedy in 1995. The thought process behind this algorithm is inspired by the social behavior of animals, such as birds in groups or groups of fish (Evanko, 2010).

Study Review

Some of the existing studies related to this study are as follows:

Table 2 - Related Research

No	Title	Author	Results	Description
1	Sentiment Analysis Of Teacher's Room App On Twitter Using Classification Algorithm	Angelina Puput Giovani, Ardiansyah Ardiansyah, Tuti Haryanti, Laela Kurniawati, Windu Gata	This study compares the NB, SVM, K-NN methods without using feature selection with the NB, SVM, K-NN methods which use feature selection and compares the Area Under Curve (AUC) values of these methods to find out the most optimal algorithm. The test results found that the best optimization application in this model was an SVM-based PSO algorithm with an accuracy value of 78.55% and an AUC of 0.853. This research managed to get the effective and best algorithm in classifying positive comments and negative comments related to the Ruang Guru application. (Giovani et al., 2020)	Jurnal Teknoinfo, Vol. 14, No. 2, 2020, 116-124, ISSN: 2615-224X DOI:10.33365 / jti.v14i2.679
2	Text Mining Accuracy Using K-Nearest Neighbor Algorithm on SMS News Content Data	Windu Gata, Purnomo	The results of the research conducted obtained results on the accuracy of the ya prediction selection of 772 correct and not in accordance with the number of 32, so that the precision was 96.02%. Meanwhile, predictions do not have a result of 0 errors and 14 correct in the prediction of NO. So that the accuracy results obtained are 96.15%.(Gata, 2017)	www.neliti.com Journal Format Volume 6 Number 1 of 2017: ISSN : 2089 -5615
3	Twitter Sentiment Analysis Of Post Natural Disasters Using Comparative Classification Algorithm Support Vector	Ainun Zumarniansyah , Rangga Pebrianto, Normah, Windu Gata	In calculating the natural disaster sentiment analysis using a comparison of the Support Vector Machine and the Naive Bayes algorithm, the difference in accuracy is 3.07% where the support vector machine results are greater than the Naive Bayes. (Zumarniansyah et al., 2020)	Jurnal Pilar Nusa Mandiri Vol 16 No 2 (2020): Publishing Period for September 2020.

No	Title	Author	Results	Description
	Machine And Naïve Bayes			https://doi.org/10.33480/pilar.v16i2.1423
4	Sentiment Analysis of Covid-19 Information using Support Vector Machine and Naïve Bayes	Ratino, Noor Hafidz, Sita Anggraeni, Windu Gata	There are several classification algorithms used, namely Naïve Bayes with an accuracy of 78.02% and an AUC of 0.714, while the Support Vector Machine produces an accuracy of 80.23% and an AUC of 0.904. It has an accuracy difference of 2.21%. After optimization with the Particle Swarm Optimization operator, the Naïve Bayes (PSO) algorithm produces an accuracy of 79.07% and an AUC of 0.729, while the Support Vector Machine (PSO) algorithm produces an accuracy of 81.16% and an AUC of 0.903. It has an accuracy difference of 2.09%. Algorithm test results, PSO-based Support Vector Machine or not, can always result in higher accuracy.(Ratino et al., 2020)	JUPITER Journal (Journal of Computer Science and Technology Research) Vol 12 No 2 (2020): JUPITER October 2020
5	Sentiment Analysis of the House of Representatives with Particle Swarm Optimization-Based Classification Algorithm	Anas Faisal, Yuris Alkhalifi, Achmad Rifai, Windu Gata	The study was conducted using two algorithms, namely the Support Vector Machine (SVM) Algorithm and Naive Bayes (NB). The two algorithms are each optimized using Particle Swarm Optimization (PSO). The results of the SVM and NB k-fold cross validation tests obtained accuracy values of 71.04% and 70.69% with Area Under the Curve (AUC) values of 0.817 and 0.661. While the results of the k-flod cross validation test using PSO, for SVM and NB, they received accuracy values of 75.03% and 73.49% respectively with AUC values of 0.808 and 0.719. The use of PSO is able to increase the accuracy value of the SVM algorithm by 3.99% and 2.8% in the NB algorithm. The result of testing the two algorithms the highest accuracy value was SVM with a PSO of 75.03%.(Faisal et al., 2020)	Jurnal JOINTECS (Journal of Information Technology and Computer Science) Vol 5, No 2 (2020) DOI: https://doi.org/10.31328/jointecs.v5i2.1362
6	Sentiment Analysis of National Exam Removal on Twitter Using Support Vector Machine and Naïve Bayes-based Particle Swarm Optimization	Yuris Alkhalifi, Windu Gata, Arfhan Prasetya, Imam Budiawan	The test was carried out using k-Fold Cross Validation to obtain accuracy values, confusion matrix tables and area under curve. The test results obtained an accuracy value of 92.92% and an AUC of 0.977 for SVM without PSO. Then the accuracy value is 94.81% and the AUC is 0.974 for SVM with PSO. The accuracy value is 85.93% and the AUC is 0.645 for NB without PSO. As well as an accuracy value of 86.92% and an AUC of 0.715 for NB with PSO. In this study, the SVM method with PSO was best for classifying positive and negative sentiments related to the elimination of UN.(Alkhalifi et al., 2020)	CoreIT Journal Vol 6, No 2 December 2020 ISSN 2460-738X (Print) ISSN 2599-3321 (Online)
7	Internet Sentiment Analysis on AMIK BSI Tegal Social Media	Ahmad Fauzi, Amin Nur Rais Muhammad Faittullah	The NAIVE BAYES algorithm and its methods will be tested with two inputs using positive (100 commentary comments) and negative (100 text comments), the accuracy obtained by the NAIVE BAYES algorithm is	Jurnal SEMNATI Vol 1 (2018): SEMNATI 2018

No	Title	Author	Results	Description
	Using Naive Bayes Algorithm	Akbar, Windu Gata	76.50%+/-7.76%(micro:76.50). The results showed that NAIVE BAYES (NB) got the best and accurate results.(Fauzi et al., 2018)	
8	Sentiment Analysis of DKI Jakarta Governor Candidate 2017 on Twitter	Ghulam Asrofi Buntoro	The data used is tweets in Indonesian with the keywords AHY, Ahok, Anies, with a total dataset of 300 tweets. The result of this study is an analysis of sentiment towards the 2017 DKI Jakarta gubernatorial candidate. The highest accuracy was obtained when using the NaÃ¯ve Bayes Classifier (NBC) classification method, with an average accuracy value of 95%, a precision value of 95%, a recall value of 95% a TP rate value of 96.8% and a TN rate value of 84.6%. (Buntoro, 2017)	Jurnal INTEGER: Journal of Information Technology Vol 2, No 1 (2017).
9	Sentiment Analysis of Online Learning on Twitter during the COVID-19 Pandemic Using the Naïve Bayes Method	Samsir, Ambiyar, Unung Verawardina, Firman Edi, Ronal Watriantho	The analysis was conducted on Twitter by mining document-based text interpreted using the Naïve Bayes algorithm. The results showed that online learning had a positive sentiment of 30 percent, a negative sentiment of 69 percent, and a neutral of 1 percent during the period. Due to public dissatisfaction about online learning, many negative sentiments were created. Some tweets show disappointment with the words 'stress' and 'lazy' in conversations that become high-frequency words. (Samsir, Ambiyar, Unung Verawardina, Firman Edi, 2021)	JOURNAL OF INFORMATICS MEDIA BUDIDARMA Vol 5, No 1 (2021). DOI: http://dx.doi.org/10.30865/mib.v5i1.2580
10	Twitter Sentiment Analysis Of Post-Covid-19 Online Lectures Using Support Vector Machine Algorithm and Naive Bayes	Hendrik Setiawan, Ema Utami, Sudarmawan	For sentiment analysis, researchers applied the Bayes nave algorithm and support vector machine (SVM) with performance results obtained on the Bayes algorithm with an accuracy of 81.20%, a time of 9.00 seconds, a recall of 79.60% and a precision of 79.40% while for the SVM algorithm it obtained an accuracy value of 85%, a time of 31.60 seconds, a recall of 84% and a precision of 83.60%, the performance results were obtained at iteration 1 for nave Bayes and the 423rd iteration for the SVM algorithm.(Setiawan et al., 2021)	Journal of Mathematics (Computing and Informatics) Vol 5 No 1 (2021) https://doi.org/10.31603/komtika.v5i1.5189

Thinking Framework

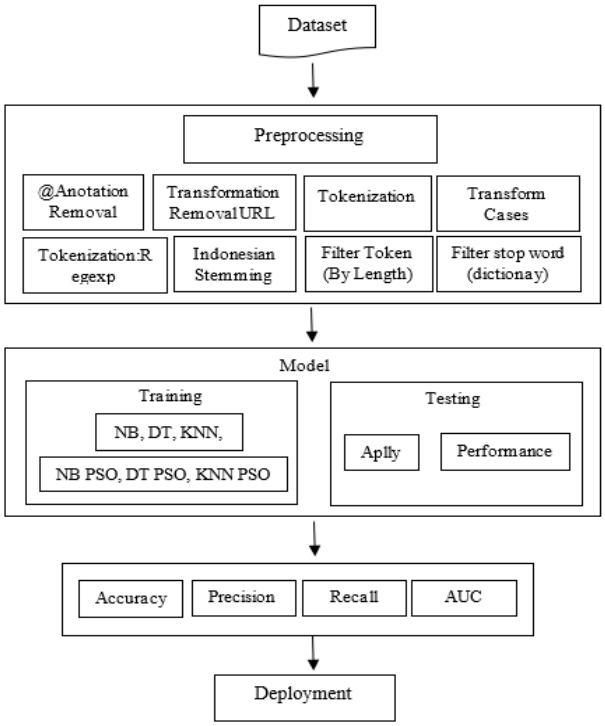

Fig. 1. Frame of Mind
Source: Research Results (2022)

3. Research Methods

Research in general can be interpreted as an effort to seek knowledge or an investigative process that is carried out actively, diligently, and systematically, which aims to find information on a particular topic. Therefore, good research methods are needed to find solutions to the problems raised. The research method that will be proposed in this study is to use the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) model. The *Cross-Industry Standard Process for Data Mining* (CRISP-DM) method consists of several 6 stages in CRISP-DM, namely *Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation* and *Deployment.*

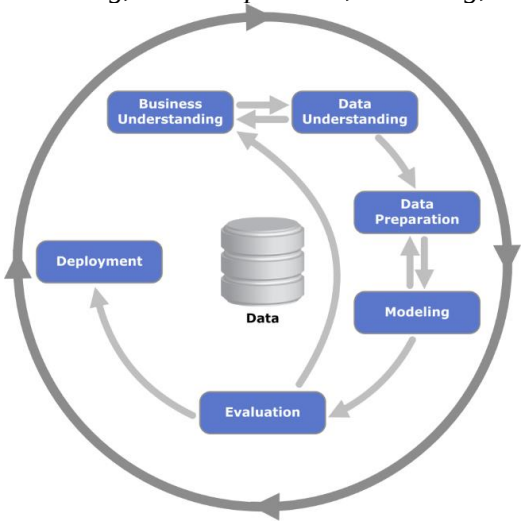

Fig. 2. CRIPS-DM Methods
Source: (Shafique & Qaiser, 2014)

4. Results and Discussions

Results and Discussion is a section that contains all scientific findings obtained as research data. This section is expected to provide a scientific explanation that can logically explain the reason for obtaining those results that are clearly described, complete, detailed, integrated, systematic, and continuous.

The discussion of the research results obtained can be presented in the form of theoretical description, both qualitatively and quantitatively. In practice, this section can be used to compare the results of the research obtained in the current research on the results of the research reported by previous researchers referred to in this study. Scientifically, the results of research obtained in the study may be new findings or improvements, affirmations, or rejection of a scientific phenomenon from previous researchers.

Business Understanding

The *Business Understanding* stage is the initial stage in research to understand the scope of the problem and determine the objectives of the research. In this study, opinions or opinions related to madrasahs from social media are very diverse, this can be used and helps to find out the public's views on madrasahs, to reveal factors that influence the results of research and produce appropriate solutions.

Data Understanding

At the data understanding stage, a process of understanding the data that will be used as research material is carried out. At this stage, the process of retrieving the original data is carried out in accordance with the required attributes. The dataset to be used is an opinion or opinion from the social media *platform Twitter*. The data collected is only *Indonesian-language tweets* from April 10, 2022 to April 16, 2022. *The* madrasa query parameter is set to 5000 and use the latest or latest type. Then save the popular file to Microsoft Excel. Using *Twitter's RapidMiner Studio Tools API version 9.10*, *Twitter's* social media crawl method is used to retrieve *tweet* data.

The data preparation stage is the stage of the data preparation process that aims to make the data clean and ready for research. The initial data obtained from *crawling* data was comments on social media *twitter* related to madrasahs as many as 3288 pieces of data. In addition, the process carried out is a *cleanup* process, such as deleting duplicate data, deleting data with narratives that are not related to the research topic, and producing 458 pieces of data.

Here's one example of *a view* or opinion on Social Media *Twitter*:

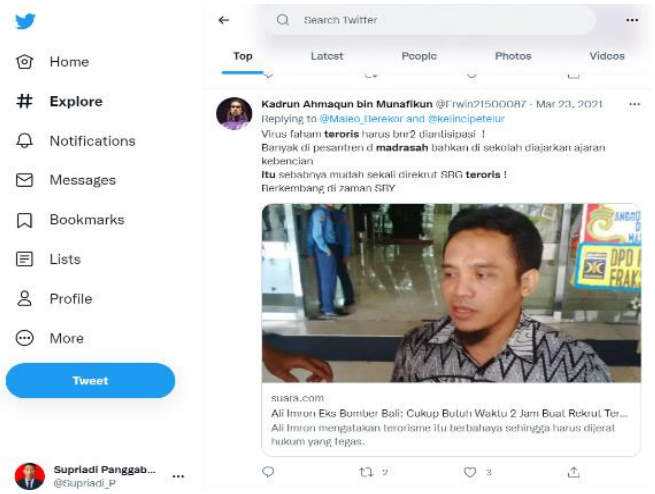

Fig. 3. Sample of Opinions on Twitter Social Media
(Source: Research Results (2022)

Using the data source obtained through the cleaning process, a data set is created with attribute text, which contains opinions or opinions narrated by the waiter that are considered consistent with the population document, then determined the class attribute. In this study, three attributes or class labels will be used in this study, namely positive and negative. In the process

of determining these attributes, it is carried out by the Madrasah Supervisor Working Group in South Jakarta.

Table 3 - Labeling Process Results Table

Labeling Process Results	
Label	Sum
Positive	233
Negative	225
Total	458

Source: Research Results (2022)

Data Preparation

The preprocessing stage is needed to clean the data from unnecessary text, where the unstructured text data will be converted into structured or semi-structured text data. The stages of preprocessing to process data are *case folding*, *convert emoticons*, *cleansing*, *tokenizing*, *stop word removal* and *stemming*.

Pthere is a first stage this researcher uses Gataframework by accessing the link *http:// www. gataframework.com/*, here's how it looks:

Fig. 4. Data Preprocessing Model Design Drawing using Gataframework
Source: Research Results (2022)

Due to system limitations, where *Gataframework* can only preprocess a maximum of 100 data, the researcher uses the *Excecute Python source code* connected from the *RapidMiner* application to *Gataframework.*

Fig. 5. Data Preprocessing using Execute Python script on Rapidminer application
Source: Research Results (2022)

Modeling

At this stage, datasets that have gone through *preprocessing* will be used as input in the classification algorithm, and used as *training* and *testing datasets.* According to the previous chapter, this study will use four Algorithms at once as comparative material, namely *Naïve Bayes (NB), Decision Tree (DT)* and *K – Nearest Neighbor (K-NN).* After the *Preprocessing* process with the Rapid Miner tool, then proceed with *Tokenization*, *Stop word Filter (Dictionary)*, *Token Filter (by Length)* and 10 *Cross fold validation*. With the 10-Fold *Cross validation* method, the dataset is divided into 10 areas, with each aspect providing the same information percentage of

each type of data. 9/10 of the data area is used in the Training process to form a model, while 1/10 of the area is used in the Testing process. Training to produce models and testing *performance*.

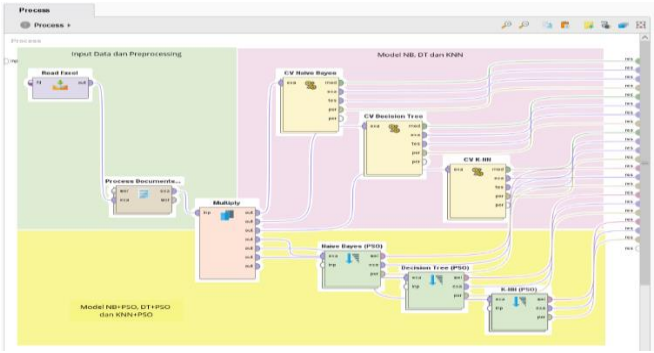

Fig. 6. Validation Testing Model Using Naïve Bayes (NB), Decision Tree (DT), K – Nearest Neighbor (K-NN) Naïve Bayes (NB) PSO, Decision Tree (DT) PSO and K – Nearest Neighbor (K-NN) PSO
Source: Research results (2022)

Evaluation

The comparison *of accuracy*, *precision*, *recall* and *AUC* between *the Naïve Bayes, Decision Tree, k-Nearest Neighbor, Naïve Bayes PSO, Decision Tree PSO,* and *k-Nearest Neighbor PSO* algorithms with *the 10-Fold Cross Validation* model has been carried out as follows:

Table 4 - Comparison of *Accuracy, Precision, Recall and AUC*

Validation	*Algorithm*	*Accuracy*	*Precision*	*Recall*	AUC
Cross Validation	NB	73.84%	79.42%	73.84%	0.712
	German	61.38%	57.21%	97.01%	0.607
	K-NN	74.70%	74.08%	78.53%	0.853
	NB PSO	80.80%	83.03%	78.64%	0.739
	DT PSO	65.27%	59.75%	98.68%	0.647
	K-NN PSO	67.24%	81.97%	52.44%	0.764

Source: Research Results (2022)

Based on the results of the comparison of research in table 4.13 from *tweet* processing as many as 458 data shows that the results of the *accuracy* pattern of classification of the *Naïve Bayes PSO* algorithm outperform other algorithms, namely *Naïve Bayes, Decision Tree, k-Nearest Neighbor, Decision Tree PSO,* and *k-Nearest Neighbor PSO*.

Table 5 - *Confusion matrix* with *Naïve Bayes PSO* model Accuracy: 80.80% +/- 4.86% (micro average: 80.79%)

	true NEGATIVE	true POSITIVE	class precision
Pred. NEGATIF	187	50	78.90%
Pred. POSITIVE	38	183	82.81%
class recall	83.11%	78.54%	

Source: Research Results (2022)

AUC: 0.739 +/- 0.105 (micro average: 0.739) (positive class: POSITIF)

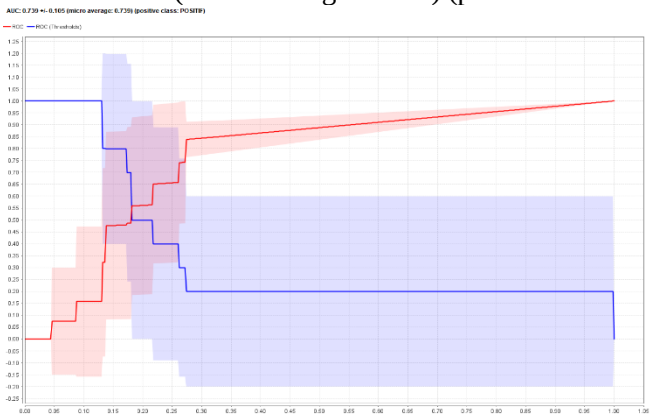

Fig. 7. AUC model *NB PSO*
Source: Research results (2022)

Based on the results of the study using the Naïve Bayes PSO algorithm, an *Accuracy*

pattern was obtained, which was 80.80%, where from 233 data predicted positive, it turned out that 183 data were correctly predicted positive (TP) while 50 data turned out to be predicted negative (FP) and from 225 data predicted negative it turned out that 187 were correctly predicted negative (TN) while 38 data turned out to be predicted positive (FN), *precision* 83.03%, *Recall* 78.64% and model performance from AUC on ROC curve is 0.739.

Deployment

Based on the evaluation results of the model testing process between the Naïve Bayes algorithm, Decision Tree, k-Nearest Neighbor, Naïve Bayes PSO, PSO Decision Tree, and k-Nearest Neighbor PSO, it was found that the highest model test results from all algorithm testing results is the Naïve Bayes PSO model. Therefore, the weights that will be used in the application modeling research are based on the results of testing the Naïve Bayes PSO algorithm.

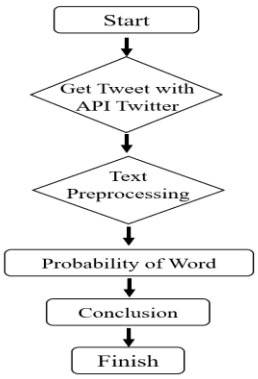

Fig. 8. Application Flowchart
Source: Research Results (2022)

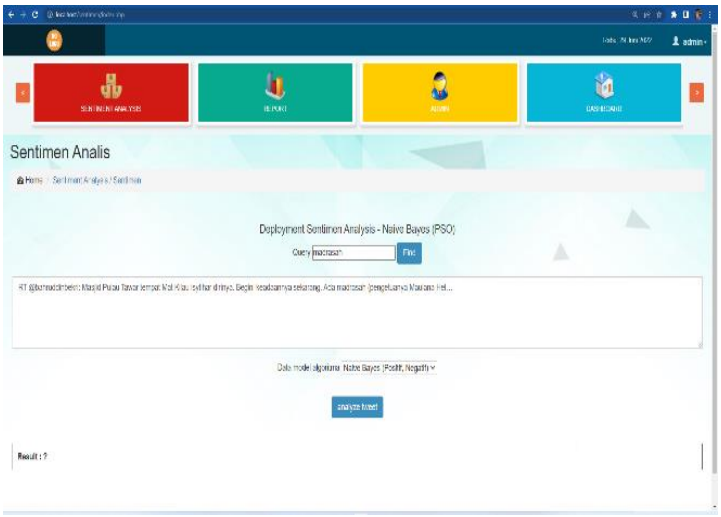

Fig 9. Get Tweet with Twitter API
Source: Research Results (2022)

The picture above shows the deployment results to get tweets to twitter using the twitter API (Application Programming Interface) by mentioning madrasahs. At this stage the tweet data is taken so that it can be carried out in the next step, namely text preprocessing.

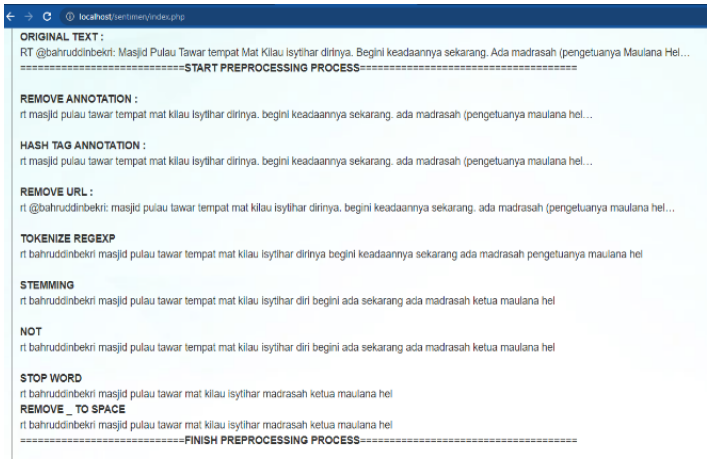

Fig. 10. Text Preprocessing
Source: Research Results (2022)

In the picture above, after the tweet data is taken, the next step is to preprocess and clean the text, using the Remove @annotation, Remove URL, Tokenize Regexp, Stemming, Not Transformation Negative, Stop word, Remove _ to Space techniques. After the tweet data has been preprocessed and cleaned, the next step is to calculate the word weight. Where the word weight is obtained from the test results of the Naïve Bayes PSO model, because Nave Bayes PSO is an algorithm that has the highest accuracy compared to the Naïve Bayes algorithm, Decision Tree, k-Nearest Neighbor, Naïve Bayes PSO, PSO Decision Tree, and k-Nearest Neighbor. PSO.

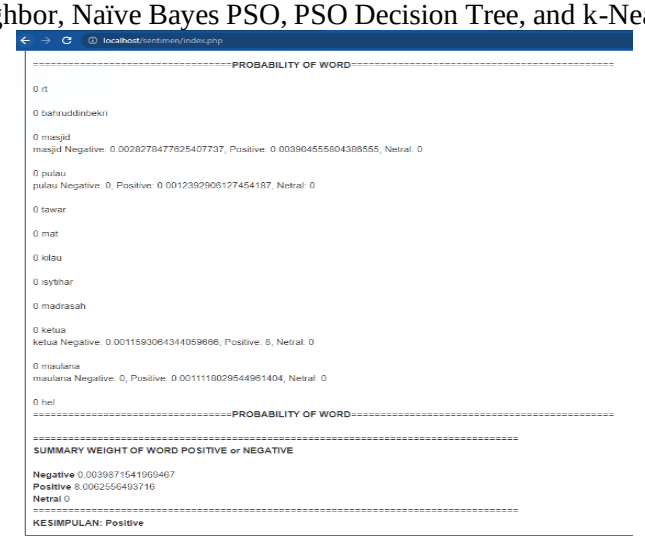

Fig. 11. Probability of word
Source: Research Results (2022)

In the picture above, it can be seen that household has a weight of 0, Baharuddin has a weight of 0, a mosque has a negative weight: 0.0028278477625407737 a positive weight: 0.003904555804386555, an island has a negative weight: 0 a positive weight: 0.0012392906127454187, a bargain has a weight of 0, a mat has a weight of 0, a shine has a weight 0, Isytihar has a weight 0, madrasa has a weight 0, the chairman has a negative weight: 0.0011593064344059666 positive weight: 8, maulana has a negative weight: 0 positive weight: 0. 0.0011118029544961404 and hel has a weight of 0. The results of the calculation of these weights are negative weights: 0.0039871541969467 and a positive weight of 8.0062556493716. Thus, the results of the calculations for these categories produce positive conclusions.

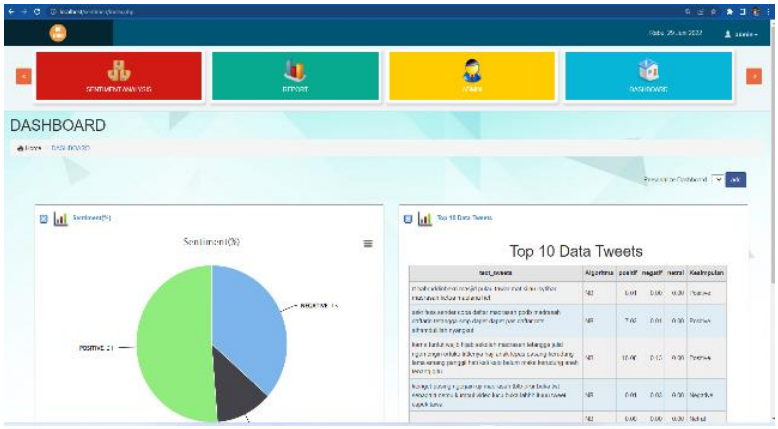

Fig. 12. Result Summary Prediction category
Source: Research Results (2022)

In the picture above, shows the graphic data and the results of the summary predictions of categories that have been categorized based on tweet data and the weight of each word in each category has been calculated. In addition, a graph function to monitor the number of tweet data that has been categorized so that the progress of the data can be monitored.

5. Conclusion

From the research above, it can be concluded that the application of Data Mining for the case of sentiment analysis towards madrasahs using 3 classification algorithms *of Naïve Bayes (NB), Decision Tree (DT)* and *K – Nearest Neighbor (K-NN).* To improve the performance of the classification method can be done using the *Particle Swarm Optimization (PSO)* selection feature. The test results of the algorithm *Naïve Bayes (NB) PSO* get the highest accuracy when compared to the algorithms *of Naïve Bayes, Decision Tree, k-Nearest Neighbor, Decision Tree PSO,* and *k-Nearest Neighbor PSO* which is 80.80%, so that with this it can be applied to analyze an opinion.

References

Aditia Rakhmat Sentiaji, A. M. B., Sarjana, P. S., Statistika, D., Matematika, F., Ilmu, D. A. N., & Alam, P. (2014). Analisis Sentimen Terhadap Acara Televisi Berdasarkan Opini Publik. *Jurnal Ilmiah Komputer Dan Informatika (KOMPUTA)*.

Alita, D., Priyanta, S., & Rokhman, N. (2019). Analysis of Emoticon and Sarcasm Effect on Sentiment Analysis of Indonesian Language on Twitter. *Journal of Information Systems Engineering and Business Intelligence*, *5*(2), 100. https://doi.org/10.20473/jisebi.5.2.100-109

Alkhalifi, Y., Gata, W., Prasetyo, A., & Budiawan, I. (2020). Analisis Sentimen Penghapusan Ujian Nasional pada Twitter Menggunakan Support Vector Machine dan Naïve Bayes berbasis Particle Swarm Optimization. *CoreIT*, *6*(2), 71–78. http://ejournal.uin-suska.ac.id/index.php/coreit/article/view/9723

Badri, M. (2011). *Corporate Marketing and Communication*. Universitas Mercu Buana.

Brahimi, B., Touahria, M., & Tari, A. (2019). Improving sentiment analysis in Arabic: A combined approach. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1242–1250. https://doi.org/10.1016/j.jksuci.2019.07.011

Brogan, C. (2011). *Social Media 101: Tactics and Tips to Develop Your Business Online*.

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *Integer Journal*, *2*(1), 32–41. https://t.co/jrvaMsgBdH

Dailey, P. R. (2009). *Social Media: Finding Its Way into Your Business Strategy and Culture*. Linkage.

Deng, L., & Yu, D. (2013). Deep learning: Methods and applications. *Foundations and Trends in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Evanko, D. (2010). Optical imaging of the native brain. *Nature Methods*, *7*(1), 34. https://doi.org/10.1038/nmeth.f.284

Faisal, A., Alkhalifi, Y., Rifai, A., & Gata, W. (2020). Analisis Sentimen Dewan Perwakilan Rakyat Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization. *JOINTECS (Journal of Information Technology and Computer Science)*, *5*(2), 61. https://doi.org/10.31328/jointecs.v5i2.1362

Fauzi, A., Rais, A. N., Akbar, M. F., & Gata, W. (2018). Analisis Sentimen Berinternet Pada Media Sosial AMIK BSI Tegal Dengan Menggunakan Algoritma Naive Bayes. *Seminar Nasional Teknologi Informasi Universitas Ibn Khaldun Bogor*, 46–54.

G.Wahyuningtyas, I. M. and S. (2014). Aplikasi Data Mining untuk Penilaian Kredit Menggunakan Metode Fuzzy Decision Tree. *Jurnal Sains Dan Seni Pomits*, *1*(1), 1–6.

Gata, W. (2017). *Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS*. *6*, 1–13.

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

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information and Management*, *52*(7), 801–812. https://doi.org/10.1016/j.im.2015.04.006

Ibrahim, D. (2017). Analisis Hubungan antar Faktor dan Komparasi Algoritma Klasifikasi pada Penentuan Penundaan Penerbangan. *2017*, *September*, 15– 17.

Kasih, P. (2019). Pemodelan Data Mining Decision Tree Dengan Classification Error Untuk Seleksi Calon Anggota Tim Paduan Suara. *Innovation in Research of Informatics (INNOVATICS)*, *2*, 63–69.

Keilany, Z. (1978). Book Reviews: Book Reviews. *Review of Social Economy*, *36*(2), 228–229. https://doi.org/10.1080/00346767800000037

Muktamar, B. A., Setiawan, N. A., & Adji, T. B. (2015). Pembobotan Korelasi Pada Naïve Bayes Classifier. *Seminar Nasional Teknologi Informasi Dan Multimedia 2015*, *2*, 43–47.

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(12), 1750–1757.

Peter Norvig, R. (2010). Artificial intelligence—a modern approach by Stuart. *Cambridge University Press*.

Pratomo, Y. (2021). Sejarah Twitter, Jejaring Sosial yang Terinspirasi dari SMS. *Tekno.Kompas.Com*. https://tekno.kompas.com/read/2021/04/14/20420077/sejarah-twitter-jejaring-sosial-yang-terinspirasi-dari-sms?page=all

Purnama, H. (2011). Media Sosial Di Era Pemasaran 3.0. Corporate and Marketing Communication. *Jakarta : Pusat Studi Komunikasi Dan Bisnis Program Pasca Sarjana Universitas Mercu Buana*, Pp 107-124.

Ratino, Hafidz, N., Anggraeni, S., & Gata, W. (2020). Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes. *Jurnal JUPITER*, *12*(2), 1–11.

Riyanto, G. P. (2021). Jumlah Pengguna Internet Indonesia 2021 Tembus 202 Juta. *Tekno.Kompas.Com*. https://tekno.kompas.com/read/2021/02/23/16100057/jumlah-pengguna-internet-indonesia-2021-tembus-202-juta

Rozi, I., Pramono, S., & Dahlan, E. (2012). Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi. *Jurnal EECCIS*, *6*(1), 37–43.

S.A Pattekari, A. P. (2012). Prediction system for heart disease using Na ̈ive Bayes. *International Journal of Advanced Com-Puter and Mathematical Sciences*, *3*(3), 290–294.

Safitri, S. I., Suhery, C., & Bahri, S. (2021). Implementasi Algoritma K–Means Untuk Clustering Sentimen Pada Opini Kualitas Pelayanan Jasa Penerbangan. *Coding Jurnal Komputer Dan Aplikasi*, *09*(02), 186–197. https://jurnal.untan.ac.id/index.php/jcskommipa/article/view/47377

Samsir, Ambiyar, Unung Verawardina, Firman Edi, R. W. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(1), 157–163.

https://doi.org/10.30865/mib.v5i1.2604

Setiawan, H., Utami, E., & Sudarmawan, S. (2021). Analisis Sentimen Twitter Kuliah Online Pasca Covid-19 Menggunakan Algoritma Support Vector Machine dan Naive Bayes. *Jurnal Komtika (Komputasi Dan Informatika)*, *5*(1), 43–51. https://doi.org/10.31603/komtika.v5i1.5189

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222. http://www.ijisr.issr-journals.org/

Suryanto, A., Alfarobi, I., Tutupoly, T. A., & Fauziahti, R. (2019). Optimasi Naive Bayes Menggunakan Optimize Weights Dan Stratified Pada Data Kredit Koperasi. *Mantik Penusa*, *3*(1), 211–219.

Turban, E. (2005). *Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia*. Andi.

Xiaojun, Z. (2011). Michael W. Berry and Jacob Kogan (eds.): Text mining: applications and theory. *Information Retrieval*, *14*(2), 208–211. https://doi.org/10.1007/s10791-010-9153-5

Y. Sunoto, B. W. (2014). Analisis Testimonial Wisatawan Menggunakan Text Mining Dengan Metode Naive Bayes DanDecision Tree, Studi Kasus Pada Hotel Hotel Di Jakarta. *Jurnal Informatika Dan Bisnis ANALISIS*, *3*(2), 39–49.

Zumarniansyah, A., Pebrianto, R., & ... (2020). Twitter Sentiment Analysis of Post Natural Disasters Using Comparative Classification Algorithm Support Vector Machine and …. *Jurnal Pilar Nusa …*, 169–174. http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/1423