Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes

Siswanto
Informatics Engineering Faculty of Information Technology
Budi Luhur University
Jakarta, Indonesia
siswantobl@gmail.com

Yuda Pratama Wibawa
Postgraduade of Computer Science
STMIK Nusa Mandiri
Jakarta, Indonesia
yudapw@gmail.com

Windu Gata
Postgraduade of Computer Science
STMIK Nusa Mandiri
Jakarta, Indonesia
windu.gata@gmail.com

Grace Gata
Information Syatems Faculty of Information Technology
Budi Luhur University
Jakarta, Indonesia
gatasmara@gmail.com

Nia Kusumawardhani
Faculty of Computer Science
Mercu Buana University
Jakarta, Indonesia
nia_wardhani@yahoo.com

***Abstract*—The high comment about the event of a motor racing motoGP race in a print media and electronic media, making the event makes the conversation of many people in the real world and in cyberspace. Especially in the digital era today is very easy for people to get the information they want, either through the website or through existing media social and sometimes the info is loaded in real time at the same time comment on the show about trending topics that exist in cyberspace. The curiosity of the public about info-info or comments circulating about the motoGP racing makes the conversation in the existing media social so that the topic becomes a popular topic in media social that post about the race of the motoGP race. This paper will do research how accurate the comments about the existing motoGP in existing media social such as twitter which became a forum for society to talk about the race of the motoGP race. In this paper will apply two classification algorithms to test how accurate the information or comments that become a lot of people talk through media social twitter. This paper will apply the Support Vector Machine and Navie Bayes algorithms in text mining processing. The result of SVM algorithm accuracy value is 95.50% while the value of NB accuracy is 93.00%.**

Keywords— Support Vector Machine Algorithm, Naives Bayes, Classification, MotoGP

I. INTRODUCTION

The high comment about the event of a motor racing motoGP race in a print media and electronic media, making the event makes the conversation of many people in the real world and in cyberspace. Especially in the digital era today is very easy for people to get the information they want, either through the website or through existing media social and sometimes the info is loaded in real time at the same time comment on the show about trending topics that exist in cyberspace.

Curiosity community about info-info or comments circulating about the motoGP racing makes the conversation in the existing media social so that the topic becomes a popular comment in social media that post about the race of the motoGP race. This paper will do research how accurate the comments about the existing motogp in existing media social such as twitter which became a forum for society to talk about the race of the motoGP race.

Twitter is an online social networking and microblogging service that allows users to post and read text messages up to 140 characters, known as chirp. Twitter was founded in March 2006 by Jack Dorsey, and its social networking site was launched in July. Since its launch, twitter has become one of the ten most visited sites on the Internet, and is dubbed with a short message from the Internet. On twitter, unregistered users can only read tweets, while registered users can write tweets via the web server interface, short messages (SMS), or through various mobile device apps.

II. RELATED WORK

Previous research related to the topic according to this writing is a study conducted by [1] with the title Calculation of Sentiment Analysis Based Comparison of Naive Bayes Algorithm and K-Nearset Neighbors-Based Swarm Optimazation Particle on Comments Incident MotoGP racer 2015. From the research resulting accuracy of 78.67% and AUC of 0.611. While the nano-based naive bayes method yielded accuracy of 82.00% and AUC of 0.681 and accuracy value ¬k-nn is accuracy of 70.67% and AUC of 0.817, then compared with kso-based pso yield accuracy value of 70, 33% and AUC at 0,500. The next study was conducted by [2] under the title of Text Mining Usage on Community Sentiment Analysis of the Changes in Basic Material Price through Twitter regarding the positive and negative responses of rising raw material prices on the market after prices rise.

Next research by[3] with the title of Trending Classification of Twitter Topics with the Application of Naive Bayes Methods to determine which topics are the most trend in media social twitter and the results obtained from the classification of trending topic using Naïve Bayes Method. The results of the test data show the results of Religion category 16.67%, Sports 36.7%, News 6.7%, Television &

Movies 6.7% and Music 33.3%. Further research conducted by [4] entitled Analysis Sentiment Prospective Governur DKI Jakarta 2017 on Twitter to determine the public opinion about the candidates for both positive, negative and neutral sentiment. The accuracy value obtained by using Naives Bayes algorithm is 95.00% and the Support Vector Machine algorithm is 90.00.

The next research conducted by[5], entitled Sentiment Analysis On Twitter Using Text Mining Sentiment Analysis On Twitter Using Text Mining denngan results reaching 93% with 2700 training data. Subsequent research conducted by [6], entitled Twitter Trending Topic Classification to determine the trending topics in social media twitter that produces a classification accuracy value of up to 65% and 70%. Further research conducted by [7] who took the title of Trending Classification of Twitter Topics with the Application of Naive Bayes Methods to determine which topics trends in social media twitter.

III. LITERATUTE SURVEY

A. Data Mining

Data mining itself according to [8] is an attempt to find interesting patterns of large amount of data, which can be stored in databases, data warehouses, or other storage places. Likewise with data consisting of tweets, the amount of data is abundant and of course has interesting patterns that can be utilized.

According to Turban et al said that "Data mining is a term used to describe the discovery of knowledge in the database. Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases [9].

Data mining is also defined as an automated process of data that is very large and aims to get a relationship or pattern that provides benefits. Data mining is also a decision support process where the search for patterns of information in the data. This search can be done by the user. This search is called discovery. Discovery is the process of searching in the database in finding hidden patterns with no previous ideas or hypotheses about existing patterns. In other words the app takes the initiative to find patterns in the data without the user thinking about the relevant question first[10].

Data mining is a process that employs one or more computer learning techniques to analyze and extract knowledge automatically. Another definition of this is induction-based learning is the process of formulating general concept definitions by observing specific examples and concepts to be studied as figure 1 [11].

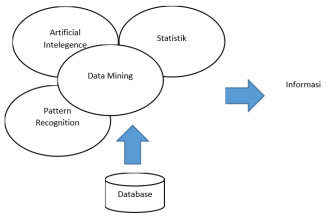

Fig 1. Root Science Data Mining

B. Document Extraction

Text that will be done text mining process, generally have characteristics such as having a high dimension, there is noise on the data, and there is a text structure that is not good. The way used in studying a text data, is to first determine the features that represent each word for each feature in the document. According to Mr. et al prior to determining the features that represent, required general preprocessing stage in text mining on the document, namely tokenizing, Filter Token, Stopword Filter and Transform Cases [12].

C. Classification

Classification is a process of data analysis that generates a model model to describe the classes contained in the data. These models are called classifier. So, this classifier will be used to arrange the classes contained in the data. There are many types of classification algorithms, two of which are Decision Tree and k Nearest Neighbor (k-NN) [13].

Classification can be defined in detail as a work that does the training / learning of the target function f which maps each vector (feature set) χ into one of a number of available y class labels. The training work will produce a model which is then stored as memory. The model in the classification has the same meaning as the black box, where there is a model that receives the input then is able to do the thinking on the input and give the answer as the output of the thought [14].

D. Text Mining

Text Mining, as one type of Data Mining is a methodology that analyzes textual data that is not easy to process algorithmically, unstructured, but is the most common form of data in the information exchange process [15].

Text mining (TM) can be defined as a scientific process by which a researcher interacts with a collection of documents using various tools to analyze the text contained in the document. The main purpose of TM is to analyze information to find patterns [16].

Text mining is a new and exciting field of research that tries to solve excess information problems using data mining techniques, machine learning, Natural Language Processing (NLP), Information Retrieval (IR), and knowledge management. Text mining involves preprocessing phases of document collections such as text categorization, information extraction, term extraction [17].

This is in line with the DM which aims to extract interesting patterns from the data, except for this data-shaped data TM for DM data is in the form of numbers [18]. DM assumes that the data is in the form of structured whereas TM unstructured data in the form of a collection of documents (corpus) should be processed first (preprocessing) into a structured form [17]. Unprocessed text usually has high dimensional characteristics, there is noise on the data and there is a bad text structure. For that, in the processing of initial data, text mining must go through several stages called prepocessing. The stages as figure 2, are:

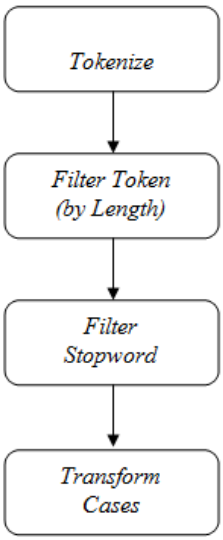

Fig 2. Document Extraction Process

1) Tokenization

Phase splits the phrase with words. By dividing first in the world, the already inputed string will be simpler because it shows in every words according to the spaces that separate it, so in a form that, will facilitate the process of changing into a word bar.

2) Filter Token (by Length)

Words that have a length of less than 4 and more than 25 will be removed, such as a word that is not, jd, ane, ga, gan, which are words that have no meaning apart if separated with other words and are not related to adjectives associated with the classification.

3) Filter Stopword

Stopwords Removal used is a stopwords filter (Dictionary) because the dataset is in Indonesian language. In this process, irrelevant words will be deleted, like words but, for, with those words which have no meaning apart if they are separated with other words and are not related to adjectives that are related to sentiment.

4) Transform cases

Converts whole letters into lowercase or all capital.

E. Support Vector Machine

Support Vector Machine (SVM) is a learning machine method that works on the principle of Structural Risk Minimization (SRM) in order to find the best hyperplane that separates the two classes in input space [19]. The best hyperplane is a hyperplane located halfway between sets of objects of two classes. The best dividing hyperplane between the two classes can be found by measuring the hyperplane's margins by finding the maximum point. Margin is the distance between the hyperplane and the nearest pattern of each class. The closest pattern is referred to as the support vector [20].

F. Naive Bayes

Naive Bayes is a simple probabilistic-based prediction technique based on Bayes's theorem (Bayes rule) with strong (naive) independence assumptions. In other words, in Naive Bayes the model used is "independent feature model". Naive Bayes is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. The performance of Naive Bayes is competitive in the classification process although it uses the assumption of attribute independence (no attribute linkage). The assumption of the dependency of these attributes on the data is rare, but although the assumptions of the dependency attribute are violated, the performance of NaiveBayes classification is quite high, as evidenced by various empirical studies [21].

The NBC method takes two stages in the process of classification of text, the training phase and the classification stage. In the training phase, the process of analyzing the sample of documents in the form of vocabulary selection, which is a word that may appear in the collection of sample documents that can be a document representation as much as possible. Next is the determination of probability for each category based on the sample document. At the classification stage, the category value of a document is determined based on the term that appears in the document classified [22].

Naive Bayes has the advantage of ease of construction and does not require complex repetition scheme parameters so it is easy to read large amounts of data. This occurs because of the design of classification guidance to the data. In addition, this method is expressed as an algorithm that has the nature of simplicity, elegance and robustness [23].

G. CRISP-DM

CRISP-DM (CRoss-Industry Standard Process for Data mining) is a consortium of companies established by the European Commission in 1996 and has been established as a standard process in data mining that can be applied in various industry sectors. The following figure describes the life cycle of data mining development that has been defined in the following CRISP-DM images as figure 3.

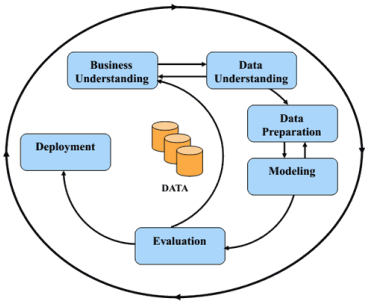

Fig 3. CRISP-DM Source: [24]

The following are six stages of CRIPS-DM development life cycle data mining:

1) Business Understanding

In this first stage it should be defined what knowledge to be gained in the form of general questions, such as how to increase profits, how to anticipate product defect errors, and so on.

2) Data Understanding

This second stage aims to collect, identify, and understand the data assets we have. The data must also be verified for truth and reliability.

3) Data Preparation

This stage includes many activities, such as clearing data, reformatting the data, reducing the amount of data, etc. aimed at preparing the data to be consistent in the required format.

4) Modeling

The model is a computational representation of the observations that are the result of searching and identification of the patterns contained in the data.

5) Evaluation

Evaluation aims to determine the value of the usability of the model we have successfully made in the previous step.

6) Deployment

Deployment is where the results of all previous stages are used in real terms.

IV. RESEARCH METHODOLOGY

A. Research Type

The two main approaches in the research are qualitative and quantitative approaches. Qualitative research methods relate to subjective judgments of attitudes, opinions, and behaviors. In general, the techniques used are interviews in certain groups and in-depth interviews. While the quantitative research method is used to examine on a particular sample, data collection using research instruments, data analysis is quantitative / statistical with the aim to test the hypothesis that has been determined [25].

B. Research Method

For Data Mining research, there has been a standard methodology called CRISP-DM or Cross-Industry Standard Process for Data Mining as figure 4.

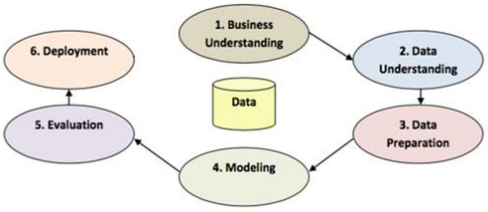

Fig 4. CRISP-DM Method

Research method used in this research is by using experiment method. This study aims to analyze the comments sentiment motoGP on media social twitter. In designing this experimental research method the researcher uses the standard research method used in data mining that is Cross-Industry Standard Process for Data Mining (CRISP-DM) that consists of six phases with the steps are Business Understanding, Data Understanding, Data Preparation, Modeling , Evaluation and Deployment.

1) Business Understanding

The first stage of CRIPS-DM is Business Understanding. In this study used a collection of text mining data obtained from media social twitter which data retrieval obtained through RapidMiner application on 28 s / d 29 October 2017. Conducted data analysis of existing text ditwitter to know a pattern of whether or not twitter with hastag studied, from the preprocesing results can be seen there are 3 references in determining the twitter is the name of the place, the name of the racer and brand motor name, to know how accurate twitter with hastag used.

2) Data Understanding Stage

The data used is data obtained from the motoGP commentary data from media social twitter, in the data can be known comments that yes motoGP and not motoGP.

3) Data Preparation Step

The amount of data obtained in this study as much as 200 records, whether it is motoGP comments and not comments motoGP, but the data still contains duplication and anomalies or inconsistent data. To get quality data, there are several preprocessing techniques used: Tokenize process, Token Filter (by Length), Stopword Filter and Transform Cases.

4) Modeling Stage

This stage is also called the stage of learning because at this stage the training data is classified by the model and then generate a number of rules. In this research, modeling using Support Vector Machine and Naïve Bayes algorithm exist in RapidMiner 8.0.

5) Evaluation phase

At this stage testing of models to obtain accurate model information. Evaluation and validation using Confision Matrix and ROC curves.

6) Deployment Stage

At this stage applied the model that has the highest accuracy or the best on the motoGP comments on the relevant media social twitter to predict whether it is true that the comment is included in the motoGP comment or not the motoGP comments by using the new data.

Here is the research method that will be done can be seen from Figure 5. research method as follows:

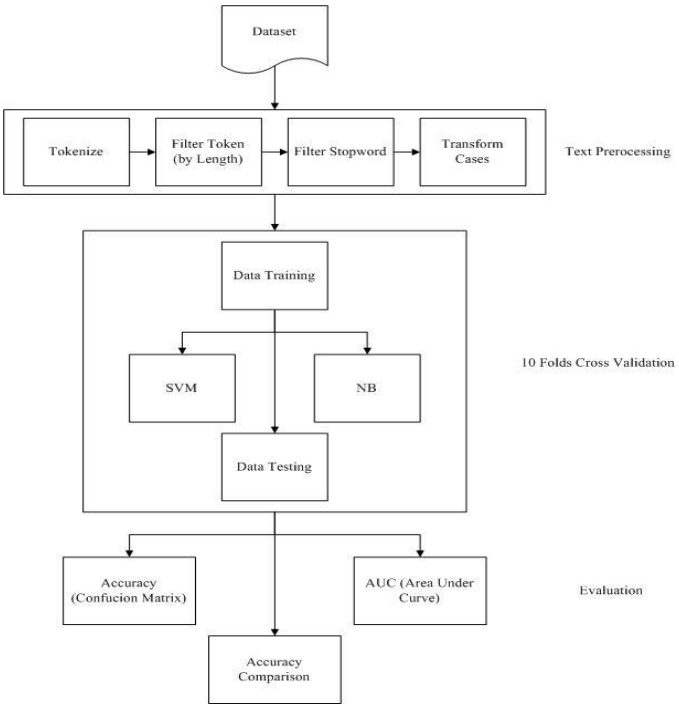

Fig 5. Research Framework

C. *Data Collection Method*

The data will be the author of the use of text data comments taken from twitter. The data the authors obtain by using RapidMiner by looking for comments about motoGP. The data collected as many as 1000 twitter text but the data taken consists of 100 comments that are motoGP yes and 100 comments that are not motoGP. Here's an example of a motoGP comment from twitter which is yes motoGP comment as tabel I and that does not comment motoGP as table II.

TABLE I. EXAMPLE COMMENTS YES MOTOGP

Content Comment	Label
26 DaniPedrosa raih pole position di MalaysianGP Sepang 2017 via btsportmotoGP MotoGP ElshintaSport	Yes

TABLE II. EXAMPLE COMMENTS NO MOTOGP

Content Comment	Label
Anda bisa memberi tanpa mengasihi. Tetapi Anda tidak mungkin mengasihi tanpa memberi. #DXPLOR.net #bunha4819	No

V. EXPERIMENTAL RESULTS

A. *Evaluation*

The experiments were performed using hardware and software specifications: Intel Core i7-4710HQ 2.5 GHz CPU, 8GB RAM, Microsoft Windows 7.1 Enterprise 64 bit. To process the data, used application RapidMiner Studio Educational version 8.1 which is a Data Mining application.

The evaluation stage aims to determine the value of the usefulness of the model that has been successfully created in the previous step. For evaluation use 10-fold cross validation. From the test results model of two algorithms used is to produce an Accuracy value (Confusion Matrix) and AUC (Area Under Curve). Then get the results of ROC graph with the value of AUC (Area Under Curve) as table III.

TABLE III. ALGORITHM TEST PERFORMANCE RESULT

Algorithm	SVM	NB
Accuracy	95,50%	93,00%
AUC	0,978	0,783

From result of comparison of performance of both algorithm above, hence result of testing of Naives Bayes higher accuracy value compared to SVM algorithm. The accuracy value for the SVM model is 95.50% and the value for Naives Bayes algorithm model is 93.00% with the difference of 2.5%.

Looking at the results of calculations in table III above by applying the classification of Area Under Curve (AUC) accuracy performance, the results of this study can be divided into two classifications, namely Good Clasification for Naives Bayes algorithm with AUC of (0.783) and Excellent Clasification for SVM algorithm with AUC of (0.978).

B. *Deployment*

To support the results of this riset writing the author has designed a simple program to know or check whether a comment in media social twitter is positive or negative and will also show the tokenize process, token filter, stopword and transform cases, as well as the results of a comment motoGP it enter positive or negative comment, there are some resistant to get results, among others as figure 6 and figure 7.

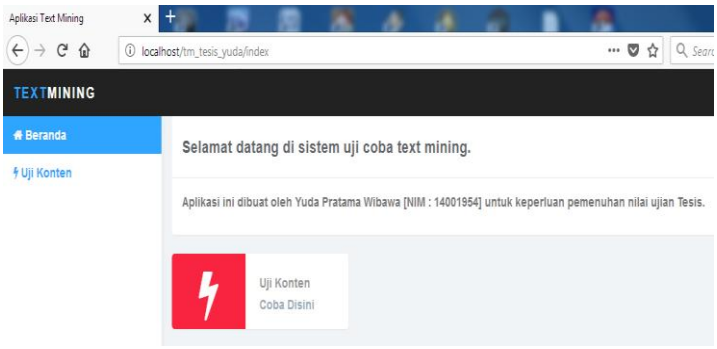

Fig 6. GUI(1)

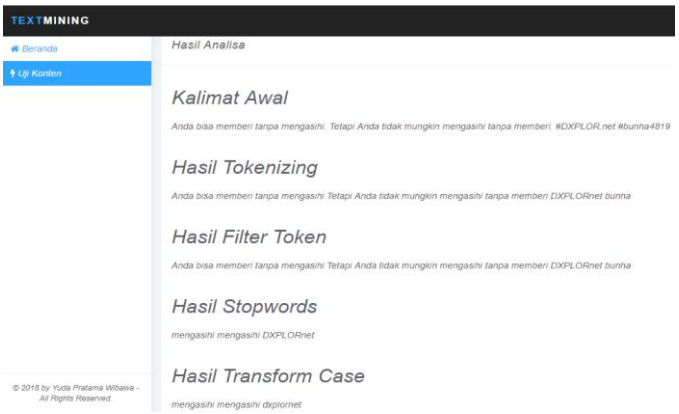

Fig 7. GUI(2)

VI. CONCLUSION

In this study after preprocessing and tested the model by comparing two methods of data mining that is support vector machine SVM and naive bayes, evaluation and validation results, it is known that the accuracy value to determine that the positive and negative comments, can be proved by the value of accuracy and value AUC of each algorithm is for SVM accuracy value = 95.50% and AUC value = 0.978, while for Naives Bayes algorithm the accuracy value = 93.00% and AUC value = 0.783.

REFERENCES

[1] Jehan. S. K., "Perhitungan Analisis Sentimen Berbasis Komparasi Algoritma Naive Bayes dan K-Nearest Neighbour Berbasis Particle Swarm Optimization pada Komentar Insiden Pembalap MotoGP", 2015.

[2] Naffisah, M. S., & Surjandari, I, "Penggunaan Text Mining pada Analisis Sentimen Masyarakat terhadap Perubahan Harga Bahan Pokok melalui Twitter", 2015, pp. 1–20.

[3] Kharde, V. A, "Sentiment Analysis of Twitter Data : A Survey of Techniques", *139*(11), 2016, pp. 5–15.

[4] Buntoro, G. A, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter", *1*(1), 2017, pp. 32–41.

[5] Lee, K., Palsetia, D., Narayanan, R., Patwary, M. A., Agrawal, A., & Choudhary, A., "Twitter Trending Topic Classification", pp. 251–258. https://doi.org/10.1109/ICDMW.2011.171, 2011.

[6] Utomo, M. S., "Implementasi Stemmer Tala pada Aplikasi Berbasis Web", *18*(1), 2013, pp. 41–45.

[7] Agustina, P. A., Matulatan, T., Tech, M., Si, M. B. S., Sc, M., Informatika, J., … Umrah, H. (n.d.), "KLASIFIKASI TRENDING TOPIC TWITTER DENGAN PENERAPAN METODE NAÏVE BAYES (The Classification Of The Trending Topic Of Twitter â€™s With Naïve Bayes Method)", 2012.

[8] Han, Kamber, Pei, Data Mining Concept and Technique. Morgan Kaufman Pubisher, 2012

[9] Kusrini, dan Emha Taufiq Luthfi, Algoritma Data mining. Yogyakarta: Andi Offset, 2009.

[10] Sari, Eka Novita, "Analisa Algoritma Apriori Untuk Menentukan Merek Pakaian Yang Paling Diminati Pada Mode Fashion Group Medan", Jurnal Pelita Informatika Budi Darma, Vol IV, No. 3, Agustus 2013, pp. 35-39.

[11] Hermawati, Fajar Astuti., Data mining,. Yogyakarta : Andi. 2013

[12] Nurhuda, Faishol, Sari Widya Sihwi dan Afrizal Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier", ISSN :2301-7201. Jakarta : Jurnal ITSMART Vol. 2 No. 2 Desember 2013.

[13] Han, Kamber, Pei, Data Mining Concept and Technique. Morgan Kaufman Pubisher, 2012.

[14] Prasetyo, Eko, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab. Yogyakarta: Andi Offset, 2014..

[15] Witten, I. H., The Practical Handbook of Internet Computing. In M. P. Singh, The Practical Handbook of Internet Computing Chapter 14 (pp. 1-23). Danvers, MA: Chapman and Hall/CRC Press. XLSTAT., 2005.

[16] Aggrawal, C., & Zhai, C.,. Mining text data. Mining Text Data (Vol. 4) http://doi.org/10.1007/978-1-4614-3223-4, 2012.

[17] Feldman R., Sanger James,The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.

[18] Weiss Sholom M., Indurkhya Nitin, Zhang Tong, Fundamentals of Predictive Text Mining. Springer-Verlag London Limited, 2010.

[19] Belloti, T., & Crook, J., "Support vector machine for credit scoring and Jiscovery of significant features. Expert System with Application},: An International Journal, 36, 2007, pp. 3302-3308.

[20] Aydin, I., Karakose, M., & Akin, E., "A multi-objective artificial immune algoritma for parameter optimization in support vector machine",. Journal Applied Soft Computing, 11, 2011, pp.120-129.

[21] Mustofa, Mufid, Pengembangan Sistem Pendukung Keputusan Penjurusan Bagi Siswa Baru Menggunakan Metode Naive Bayes. Polteknik Negeri Malang, 2016.

[22] Hamzah, Amir, "Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) Untuk Pengelompokan Text Berita dan Abstrak Akademis", ISSN:1979-911X. Yogyakarta : Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, 2012.

[23] Subiyakto, A'ang, Penggunaan Algoritma Klasifikasi Dalam Data mining. Jakarta : Syarif Hidayatullah State Islamic University, 2008.

[24] North, Matthew, Data Mining for The Masses. A Global Text Project Book, 2012.

[25] Sugiyono, Metode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung:Alfabeta, 2013