Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
13/2 (2020), 109-122. DOI: http://dx:doi:org/10:21609/jiki:v13i2.885

SENTIMENT ANALYSIS ON E-SPORTS FOR EDUCATION CURRICULUM USING NAIVE BAYES AND SUPPORT VECTOR MACHINE

Rian Ardianto, Tri Rivanie, Yuris Alkhalifi, Fitra Septia Nugraha, Windu Gata

Faculty of Computer Science, STMIK Nusa Mandiri, Jakarta, Indonesia

E-mail: 14002391@nusamandiri.ac.id

Abstract

The development of e-sports education is not just playing games, but about start making, development, marketing, research and other forms education aimed at training skills and providing knowledge in fostering character. The opinions expressed by the public can take form support, criticism and input. Very large volume of comments need to be analyzed accurately in order separate positive and negative sentiments. This research was conducted to measure opinions or separate positive and negative sentiments towards e-sports education, so that valuable information can be sought from social media. Data used in this study was obtained by crawling on social media Twitter. This study uses a classification algorithm, Naïve Bayes and Support Vector Machine. Comparison two algorithms produces predictions obtained that the Naïve Bayes algorithm with SMOTE gets accuracy value 70.32%, and AUC value 0.954. While Support Vector Machine with SMOTE gets accuracy value 66.92% and AUC value 0.832. From these results can be concluded that Naïve Bayes algorithm has a higher accuracy compared to Support Vector Machine algorithm, it can be seen that the accuracy difference between naïve Bayes and the vector machine support is 3.4%. Naïve Bayes algorithm can thus better predict the achievement of e-sports for students' learning curriculum.

Keywords: *text mining, sentiment analysis, naïve bayes, support vector machine, SMOTE*

Abstrak

Perkembangan pendidikan e-sports tidak sekedar bermain game, tetapi mengenai seluk beluk game dari mulai pembuatan, pengembangan, pemasaran, penelitian serta bentuk edukasi lain yang bertujuan melatih skill dan memberikan pengetahuan dalam membina karakter. Pendapat yang disampaikan oleh masyarakat dapat berbentuk dukungan, kritik dan masukan. Komentar dengan volume yang sangat banyak perlu dianalisis secara akurat agar dapat dipisahkan antara sentimen yang positif dan sentimen yang negatif. Penelitian ini dilakukan untuk mengukur pendapat atau memisahkan antara sentimen positif dan sentimen negatif terhadap pendidikan e-sports, sehingga dapat dicari informasi yang berharga dari media sosial. Data yang digunakan dalam penelitian ini didapatkan dengan melakukan crawling pada media sosial twitter. Penelitian ini menggunakan algoritma klasifikasi yaitu Naïve Bayes dan Support Vector Machine. Perbandingan dua algoritma menghasilkan prediksi yang diperoleh bahwa algoritma Naïve Bayes dengan SMOTE mendapatkan nilai akurasi 70.32%, dan nilai AUC 0.954. Sedangkan Support Vector Machine dengan SMOTE mendapatkan nilai akurasi 66.92% dan nilai AUC 0.832. Dari hasil ini dapat disimpulkan bahwa algoritma Naïve Bayes memiliki akurasi yang lebih tinggi dibandingkan dengan algoritma Support Vector Machine, terlihat perbedaan akurasi antara naïve bayes dengan support vector machine 3.4%. Algoritma Naïve Bayes dengan demikian dapat memprediksi pencapaian prestasi e-sports untuk kurikulum pendidikan belajar siswa dengan lebih baik.

Kata Kunci: *text mining,* analisis sentimen, *naïve bayes, support vector machine, SMOTE*

1. Introduction

The entertainment industry has always been the center of attention of all parties and is growing rapidly, it is important for all those who consider the entertainment industry to have a very high audience. Nowadays technological developments are no longer focused on content in the broadcast industry such as film, music, stage drama and others. The video game industry is also something that deserves attention, seeing its growing demand and many game application developers who follow a fairly large trend for various groups. Video games were originally only considered a hobby and often ignored, now of all ages and genders around the world playing video games and has developed into a promising income industry [1].

However, from various expectations, it is desirable to have a good perspective, assuming it is always positive, but in reality the activities of all gamers have many positive and negative points of view from the impact of playing games. One of the negative effects that often occur among gamers without supervision from parents can cause addiction from curiosity that is owned by the player, thus causing dependence to play it again [2].

In a study conducted by Syahran on several conflict addiction games found the results of research from a video game expert at Nowingham Trent University in America, Mark Griffiths, almost everyone's daily activities are playing games from all walks of life. Psychology expert in America, David Greenfield, found that about 6% of internet users experience online game addiction, which is more worrying about 7% of playing time at least 30 hours / week, this is because the average child aged 12-18 years often play games online by frequently browsing the internet that is not protected from bad information [3].

Minister of Education and Culture Regulation No. 62 of 2014 concerning Extracurricular Activities to support the achievement of educational goals in Basic Education and Secondary Education, stated that extracurricular activities carried out by students outside of hours of learning, intracurricular activities and curricular activities, under the guidance and supervision of education units, aim to develop potential, talents, interests, abilities, personality, collaboration, and independence of learners optimally [4].

Extracurricular has many forms provided in each school based on the interests and talents of students, such as Flag Raisers (PASKIBRA), Youth Red Cross (PMR), SCOUTS, Mosque Youth Association (IREMA), several art activities such as Modern Dance and Traditional, Choir, Marcing Band, sports activities such as Badminton, Soccer, Futsal, Volleyball, including electronic sports called E-sports (Electronic Sports). Understanding e-sports in general is a branch of virtual sports so that players are not directly involved in which aspects of the sport are facilitated by electronic systems, and carried out online so that each team can compete without face to face [5].

E-sports games are developing rapidly in the field of science and technology and become a determining factor in the world of education. In the current generation, students are very responsive in responding to technological developments, because many online games require players to use good skills in managing strategies, managing teamwork, negotiating and how to make the right decisions, for example online games that are included in e-sports events such as DOTA, Arena of Valor, Free Fire, Point Blank, Mobile Legend, PUBG, PES Soccer, and many more [6].

When juxtaposed with each other in terms of education and online gaming there are indeed many pros and cons. The counter statement in society is that the stigma of children playing games too often makes them forget about time and often ignores the priorities of learning in their education. With this in mind that the character crisis of children caused by technological advances is increasingly easy to access and all digital automatic. Changes in children's play activities that have a lot of interest in modern games (digital) are increasingly acute so that it greatly affects the behavior and habits of children. The phenomenon that emerges is very alarming, affecting children's learning achievement, character crisis and having aggressive behavior, even plunging children into criminal acts that can lead to death [7].

Another opinion says that e-sports does not mean that it is only applied for entertainment, but there are values that have education in students. E-sports will become a useful activity and rich in value when the activity is carried out in a directed and continuous manner, of course with the guidance and supervision from various related parties such as parents and the school [8].

Factors of the problem of addiction playing games greatly affect the youth of students, resulting in a decrease in the level of concentration in the learning process. The e-sports education carried out must have a clear purpose. If it is just entertainment, losing will not be a problem. But it must be accompanied by the aim of becoming a player who is serious about exercising e-sports in the sense of being a professional like a true athlete. Practicing with certain rules, systems, strategic patterns and various supporting aspects for the sustainability of e-sports itself is conducive [9].

The development of e-sports education is not just playing games, but also provides a basic understanding of the ins and outs of the game from the start of making, developing, marketing the e-sports game business, research and other forms of education that can support the interests and talents of students and aim to practice skills and provide adequate knowledge so that students who participate in this activity can have the necessary foundation in fostering character [6].

Various methods are used to find out what causes the addiction of online games to learning achievement. Therefore, with the rapid development of technology and knowledge-based computer systems, it has become part of the study that researchers must be involved in every field of computer science. This research was conducted to help solve problems by using data mining classifications to find out the predictions of e-sports achievements for student learning curriculum. The need for a method that can process data has been collected from the results of data collection conducted in this study.

In previous studies about the influence of online games for the prediction of learning achievement using the naïve Bayes algorithm, random forest, and C4.5 [2]. Problems with aggressive behavior of adolescents in Samarinda explain that changes in behavior are caused by online games played by teenage students [10]. The relationship of playing games with the motivation of middle school students in the Bacolod West district is a very significant relationship because online games can make the concentration of learning disrupted [11]. The results of online game addiction research in Indonesia show that middle school students spend a lot of time playing online games [12]. Online game addiction triggers an impact on attraction, aggressiveness, and interpersonal relationship problems that lead to psychological [13]. In other literature which explains that online games are not considered to have a negative impact [14]. Online games make Junior High School 1 Kuta students experience a decline in achievement [15]. The question arises, whether there is a relationship of active promiscuity, parental guidance, and discipline to learning achievement [16].

This research is increasingly interesting with a combination of education and computer science, namely data mining. Some related studies include: Handling Unbalanced Data in Predicting Churn Customers Using Combined Sampling and Weighted Random Forest [17]. Predict the time span of how long students can actively learn with the Random Forest method. Hypertension Prediction System Using Naive Bayes Classifier [18]. Nonlinear Methodology for Identifying Seismic Events and Nuclear Explosions Using Random Forests, Support Vector Machines, and Naive Bayes Classifications [19]. Prediction of Timeliness of Graduation Students Using Naïve Bayes: Case Study at Syarif Hidayatullah Islamic State University Jakarta [20]. Student Academic Performance Evaluation Using the Naïve Bayes Algorithm (Case Study: Fasilkom Unilak) [21]. Naive Bayes Method for Graduation Prediction (Case Study: New Student Data) [22]. Data mining to predict the type of transaction in cooperative loans with the C4.5 algorithm [23]. Decision Tree-based decision support system in providing scholarship case studies: AMIK "BSI yogyakarta" [24]. Sentiment analysis of public opinion on forest fire news through comparison of Support Vector Machine algorithm and k-nearest neighbor based particle swarm optimization [25]. Application of C4.5 algorithm based on particle swarm optimization for ease of service programs predicted results of donations tithes [26]. Implementation of Naïve Bayes Algorithm, Random Forest. C4.5 about Online Games for Prediction of Learning Achievement with MAN 4 Karawang student research objects [2].

The data on the website and social media is very much, so it is very difficult to detect sentiment [32]. Twitter users create their own words and by using spelling and punctuation, making misspellings, using slang words, new words, adding url, and special terms and abbreviations according to their age classification. Thus, such texts demand to be corrected. So to analyze the characters of HTML text, slang words, emoticons, words that have no meaning, punctuation, and url need to be deleted [33].

Data on Twitter social media becomes very interesting material to be analyzed because of several things including, 1) most e-sports players have an account and are active on social media, 2) many e-sports tournament information that publicly convey their ideas and ideas as well provide comments on policies issued by related parties, 3) the community freely submits ideas, support, responses and criticisms to the government regarding e-sports policies in the world of education.

Opinion Mining (OM) and Sentiment Analysis (SA) are two emerging fields that aim to help users find opinion information and detect sentiment polarity. OM and SA are generally used interchangeably to express the same meaning. However, some researchers state that they aim to overcome two problems slightly different [34]. Sentiment Analysis builds a system that tries to identify and extract opinions in texts [5]. This highlights the classification of texts and relates to the extraction of texts. Usually sentiment polarity is classified as positive, negative or neutral class [36].

Sentiment can be found in various mass media lines, such as: Facebook, tweeters or comments on a product to provide useful indicators for various purposes. It also states that a sentiment can be categorized into two groups, namely negative and positive words. Sentiment analysis is a natural language processing technique for measuring opinions or sentiments expressed in tweet choices [37].

Sentiment analysis with Twitter has recently become a popular method for organizations and individuals to monitor public opinion of their brands and businesses. One of the main challenges that must be faced by Twitter sentiment analysis method is the noisy nature of the data generated by Twitter. Twitter only allows for 140 characters in each post, which affects the use of abbreviations, irregular expressions, and rare words. This phenomenon increases the level of data sparsity, which affects the performance of Twitter sentiment classifiers [34]. The well-known method for reducing textual data noise is the elimination of stop words. This method is based on the idea that discarding non-discriminatory words reduces the classifying feature space will get more accurate results [38]. Therefore, it is necessary to have a pre-processing process that aims to process text into a standard form so that it is easy to process further. The results of the process of pre-processing is a basic word that is not a collection of stop words. Stop words in sentiment analysis are words that have weak semantics and sentiments in the context [39].

Twitter Sentiment Analysis can be one powerful tool for analyzing valuable reflections from public perception. Twitter sentiment analysis has attracted a lot of attention because of the rapid growth in Twitter's popularity as a platform for people to express their opinions and attitudes towards various topics. Twitter's sentiment analysis approach tends to focus on identifying individual tweet sentiments [40].

Sentiment analysis is the process of extracting and processing data automatically using certain algorithms to get sentiment information contained in an opinion sentence [41]. Sentiment analysis refers to a general method for extracting polarity and subjectivity from semantic orientation which refers to the strength of words and text or polarity phrases [42]. Analysis of social media data can help businesses, governments, security organizations and the environment to find out people's problems, suggestions and criticisms and find the right solutions for their problems. Analyzing social media data needs to adapt to new methods and tools, it also needs a better understanding of people's opinions and their criticisms and insights. Sentiment analysis is a field of research that emerged from Natural Language Processing (NLP) to extract people's opinions, thoughts, and views [43].

Among supervision machines, SVM is a popular learning machine that can learn from training data and classify vectors in features into one or two groups [29]. Vector Machine Support determines the linear lines (hyperplane) that separate data into categories by calculating the longest distance between support vector categories (data) [30]. NB is a learning algorithm that is based on Bayes theory using strong assumptions. Bayes theory is a theory about finding the highest probability of something based on existing data [31].

This research was conducted by applying the data mining method to compare with the Naive Bayes method and support vector machine to find the algorithm that has the highest accuracy, in terms of predicting online game addiction about the achievement of e-sports for students' educational curriculum which has never been done in research.

2. Methods

In this paper CRISP-DM stands for Cross-Industry Standard Process Model for Data Mining. Generally explained about the data mining process in six stages of Business Understanding, Understanding, Data Preparation, Data Modeling, Evaluation, Deployment, can be seen in Figure 1.

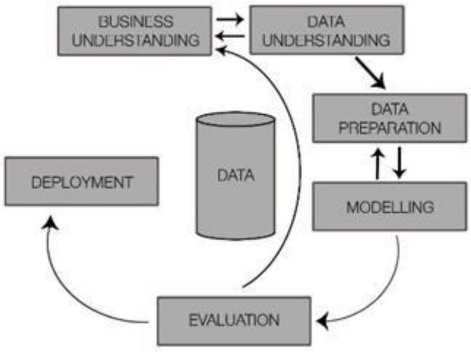

Figure 1. CRISP-DM [27]

2.1 Business Understanding

In this study, it is part of understanding the research topic that is being carried out at the stage of understanding the work. In this research, the subject of the research is understood by digging information on social media Twitter using individual electronic mathematical key structures, graphics on the fast-point vehicle to hang the object in tweets. The motivation at this point is that the tweets provided are usually in the form of text in computerized media, grouped according to the content of the discussion for each comment category. Online media is not only a way to read the main page of the article, it can also be used to see the problems that arise and even to see the training options. This sentiment analysis is done to find a classification method that can help identify positive and negative news article comments. At this point, it is understood that finding the best classification method can help in processing the information to be made by comparing the results of the algorithm used and improving the performance of the classification method that can be made using the selected features.

2.2 Data Understanding

In the next step, the raw extraction formulation is carried out according to the required characteristics. Through the main structure dedicated to e-sports, information is obtained from social networking sites on Twitter. Information was collected from 2 May 2020 to 11 May 2020. Initial information obtained was 15,000 comments Information from Tweet Comments. After gathering information, a cleanup group was conducted, because the information was still random and there was some information that was not appropriate for the research content, specific information was only suitable for sports content as a training curriculum, and an information survey was carried out using the Ms. program. Expecting Expectations 2010 uses a combination (uppercase / lowercase format) to delete duplicate information, differentiate between uppercase and uppercase letters, and then clean with a faster program with the document annotation tool. General information obtained later was 8,453 comments. This search uses Indonesian commentary information.

2.1 Data Preparation

The information preparation phase is a combination phase of information preparation aimed at obtaining information that is clean and ready for use in research. In the initial stages of text extraction, the introductory text phase will be applied; at this point, the researcher will use the RapidMiner tool. At this point, the researcher will create many preprocessing text structures in the commentary dataset, including case conversions, symbols, long filtering filters, keyword channels. The discussion of these stages will be explained in more detail in the next section.

2.4 Modelling

It is the stage of selecting data mining techniques by determining the algorithm to be used. This research uses tools that are used to do modeling in accordance with predetermined techniques, these tools are RapidMiner version 8.2. This study uses 2 classification algorithms as its model [28]. The classification algorithm used is Naïve Bayes (NB) and Support Vector Machine (SVM). The test results for each model are to categorize positive tweet articles and negative tweet articles to achieve the best accuracy value in each algorithm.

2.5 Evaluation

The evaluation phase aims to determine the usefulness of the model that was successfully created in the previous modeling step. In this study, the evaluation phase is used in conjunction with the Synthetic Minority Over-Sampling Technique (SMOTE). RapidMiner version 8.2 is used to help compare SMOTE, vector drive algorithms and artificial minority algorithms and support machines to find two different grouping methods between Naïve Bayes algorithms and data sets and compare them with Naïve Bayes algorithms. In this study, the purpose of using the SMOTE technique is to increase the accuracy value of the results obtained from the algorithm method, and thus to find the best algorithm method for this research [28].

2.6 Deployment

The deployment stage is the stage used to create an implementation model that is created in a tool that can be built with various types of programming. Making this implementation model uses the results of the experimental and evaluation process as a source of reference data. Deployment used in this research is the implementation phase of the results of the comparison of 2 algorithms, then of the 2 algorithms that have the highest accuracy value can be used as a development material in predictions of online game addiction about the achievement of e-sports for students' learning curriculum.

2.7 Weighting Word

Property, called the word weight or weighting property, is a combination that evaluates each attribute based on its relevance and its impact on the classification results [31]. This value can then be used as a basis for determining features based on the lowest weight calculated from each feature. This is weighted using the TF-IDF (Term Frequency - Inversion Document Frequency) method. The TF-IDF algorithm is one of the algorithms in the text extraction weighting feature. TF is a repetition of terminology in additional documents. The higher the number of terms (high TF) in the document, the higher the weight or the higher the match value. The type of TF equation commonly used in calculations is pure TF (raw TF). Pure TF (raw TF), the TF value is given depending on how often the term appears in the document. For example, if it happens five (5) times, the individual structure will be a value of five (5). IDF (Inversion Document Frequency) is an account term that is commonly distributed in registered documents. IDF shows relationships as terms in the document. The lower the number of documents containing the required requirements, the more valuable the IDF. The frequency of inverse documents (IDF) is calculated using equation, can be seen in equation (1) [31].

$$IDF_j = log\left(D/df_j\right) \qquad (1)$$

In equation (1) where D is the number of all documents in the collection while df_j is the number of documents containing the term (TF) [31].

$$TFIDF = TF\ x\ IDF \qquad (2)$$

In equation (2) Thus the general formula for TF-IDF Term Weighting is a combination of the standard TF calculation formula with the IDF formula by multiplying the TF value with the IDF value.

2.8 Synthetic Minority Over-Sampling Technique

In previous studies Synthetic Minority Over-Sampling Technique (SMOTE) balanced the dataset by synthesizing minority data synthetically in the input space based on their environmental information. The training data set consists of minority data points (S_{min}) and majority data points (S_{maj}). For each (X_i, Y_i) ∈ S_{min}, most data points are set (S_{maj}) [28].

For each (X_i, Y_i) ∈ S_{min}, SMOTE generates a new minority data point along the joining line segment (X_i) and one of the closest neighbors chosen at random. SMOTE can be seen in equation (3) [28].

$$X_{new} = X_i + \left(X_j/X_i\right) x\ \delta \qquad (1)$$

3. Results and Analysis

This research uses data taken from tweets comments on Twitter social media related to e-sports as mentioned in the understanding data section above. The data taken as a whole is 15000 comment data presented in Figure 2. Then the data will be done in the initial stages of data cleansing, data cleaning is done using Ms. software. Excel 2010 using the process of removing duplicate data, to remove between capital letters and non-capital letters can be distinguished (match case), and then cleaning is performed using RapidMiner software using the process document from data tool which then makes all small letters (transform cases use lower case), delete words that are less than 3 letters (filter token by length), then all words and symbol characters or special characters that are not needed in each document are collected and removed such as then specify the username mentioned (username), delete hashtag (#hastag), delete punctuation, delete numbers because only text data is used, delete link (http: //), delete special characters such as symbols or expressions, delete words foreign words Only need Indonesian words (tokenizing use regular expressions) and are administered to the label (class) manually with the help of Indonesian language experts using techniques using multiple labels to label large data (crowdsourced labeling) with positive or negative labels on each comment. From the initial stage, 8453 comments of data were labeled, so the data will be the dataset in this study.

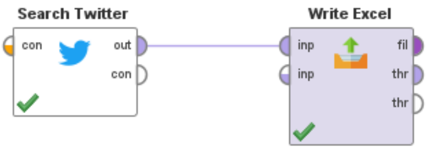

Figure 2. Process Crawling Data Twitter

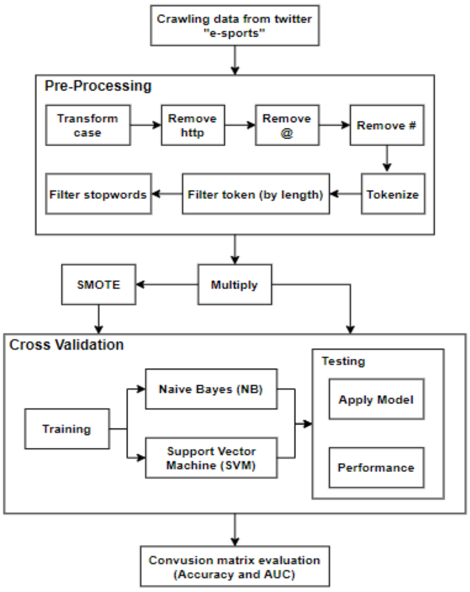

Figure 3. Sentiment Analysis Framework

In processing the data to get a model that fits the case of this research, namely, the sentiment analysis of online game addiction predictions about the achievement of e-sports for student learning curriculum using the Naïve Bayes classification algorithm and Support Vector Machine, a RapidMiner tool version 8.2 is used. Because this research belongs to the part of text-mining, there will be a stage that must be done first before a good model can be found in the case study of e-sports analysis sentiments for the education curriculum. So that the research carried out can be traced and can be re-tested, the steps to be carried out in this study are outlined in a research framework model. The stages in this framework model will be used as a reference during the research process. The framework model in this study is presented in Figure 3.

3.1 Pre-processing

The discussion at this stage is the initial process of processing the dataset before it can be processed for classification using the Naïve Bayes algorithm (NB) and Support Vector Machine (SVM) with Synthetic Minority Over-Sampling Technique (SMOTE). This study uses several stages of pre-processing for the comment text dataset, the following are the steps in Figure 4.

3.2 Transform Case

At this stage Transform Case on RapidMiner. This is used to convert all capital words to lowercase letters. The results of the conversion status gallery can be seen in Table 1.

TABLE 1
EXAMPLE OF TRANSFORM CASE

Before	After
RT @oreoqueenos: Di kelas kalian ada game apa aja ? Di kelas ku ada uno kartu , Uno stacko , moba , pubg sama scrabble ?? yang mau liat list juara sekolah kita disini https://games.grid.id/read/151616294/inilah-para-pemenang-kompetisi-mobile-legends-di-turnamen-next-2019?page=all	rt @oreoqueenos: di kelas kalian ada game apa aja ? di kelas ku ada uno kartu , uno stacko , moba , pubg sama scrabble ?? yang mau liat list juara sekolah kita disini https://games.grid.id/read/151616294/inilah-para-pemenang-kompetisi-mobile-legends-di-turnamen-next-2019?page=all

3.3 Tokenizing

At this point, continue with the combination of RapidMiner Tokenize. This is used in such a way that all special structures are independent structures and mention usernames (usernames), delete hashtags (#hastag), delete punctuation marks and delete unnecessary symbols or special characters in each document. Only delete special characters, such as text used, delete links (http: /), emoticons or emojis, delete individual structures. Foreign structures are only assigned because they take on individual organizations in Indonesia. For individual organizations, institutions designated with "no" in the future will be normalized using the underscore "_" to achieve a clear meaning, for example, "not assigned" means "not good" because the designated structure does not mean that the assigned structure is negative. Can be seen in Table 2.

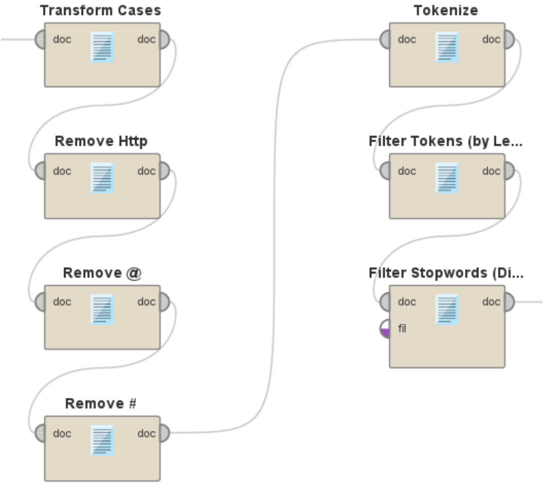

Figure 4. The stage in Pre-processing

TABLE 2
EXAMPLE OF TOKENIZING

Before	After
rt @oreoqueenos: di kelas kalian ada game apa aja ? di kelas ku ada uno kartu , uno stacko , moba , pubg sama scrabble ?? yang mau liat list juara sekolah kita disini https://games.grid.id/read/151616294/inilah-para-pemenang-kompetisi-mobile-legends-di-turnamen-next-2019?page=all	di kelas kalian ada game apa aja di kelas ku ada uno kartu uno stacko moba pubg sama scrabble yang mau liat list juara sekolah kita disini gamesgrididread151616294inilahparapemenangkompetisimobilelegendsditurnamennext2019pageall

3.4 Filter Token by Length and Filter Stop words

The crawling stage is the individual creation phase for the custom organization of token results. The stop list algorithm can be used (the least important individual institution can be deleted) or the list of individual organizations (can save each of the organizations of interest). At this point, the individual structure of the customs company used (in Indonesia) must be normalized to a standard form, for example "ke" to. The results of the filter formulation can be seen in the Table 3.

TABLE 3
EXAMPLE OF FILTERING

Before	After
di kelas kalian ada game apa aja	kelas kalian ada game apa aja
di kelas ku ada uno kartu uno stacko moba pubg sama scrabble	kelas ada uno kartu uno stacko moba pubg sama scrabble
yang mau liat list juara sekolah kita disini gamesgrididread15161629 4inilahparapemenangkomp etisimobilelegendsditurna mennext2019pageall	yang mau liat list juara sekolah kita disini

3.5 Model Classification

The next step in this research is to make a model using a classification algorithm for the comment text dataset that has gone through the pre-processing stage. This stage uses two classification algorithms, namely Naïve Bayes (NB) and Support Vector Machine (SVM) with Synthetic Minority Over-Sampling Technique (SMOTE). This study uses the RapidMiner version 8.2 tool to process the comment text dataset that has gone through the data preparation stage with text pre-processing.

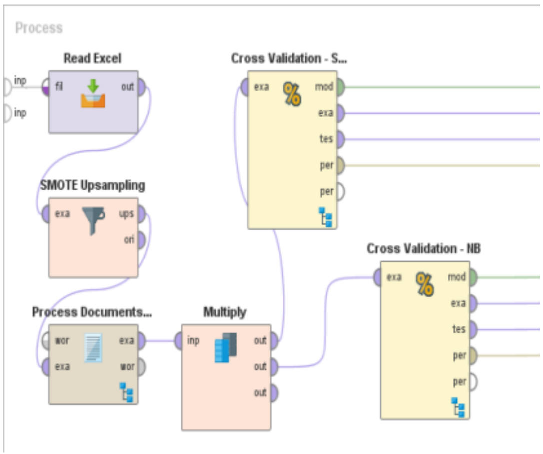

Figure 5. Modeling

The first stage of this process is that the comment text data will be uploaded into the tool by using an excel file which will then be processed with the Naïve Bayes algorithm (NB) and Support Vector Machine (SVM) to get the initial results of each algorithm, such as can be seen in Figure 2. After the first stage is carried out, this study continues by comparing the two algorithms by adding the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm. The step of using SMOTE in the modeling process aims to increase the value of the accuracy of the classification results of NB and SVM algorithms, the process can be seen in Figure 5.

3.6 Evaluation Model Classification

The evaluation phase aims to determine the ease of use of the model that was successfully created in the previous step. 10 times cross validation is used for evaluation.

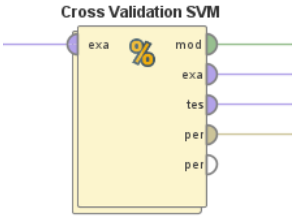

Figure 6. K-Fold Cross Validation

In the test shown in Figure 6, this number is arranged in 10. Therefore, the data set is divided into 10 regions, and each direction provides the same percentage of information for each type of information. The information used is clean and pre-made information. This information is taken from the Read Excel manager, and this is done because the dataset is stored in Excel. Handling documents from notes to convert documents into documents. Verification data consists of training information and testing information. At this point, SMOTE is used. Destructive sampling is used to balance information. For managers, "cross validation" is used to classify and evaluate agitation analysis through a 10-fold verification experiment.

TABLE 4
CONFUSION MATRIX NB WITH SMOTE

	Before	After
TP	418	425
FP	160	78
TN	196	776
FN	436	429

Based on the research data using the heat load matrix in Table 4 and Table 5, the value of accuracy, precision and recall made by SMOTE Sampling is shown can be seen in Table 6 and Table 7.

As for the comparison of accuracy and curve (AUC) regions, the results of the algorithm used can be seen in Table 8.

TABLE 5
CONFUSION MATRIX SVM WITH SMOTE

	Before	After
TP	840	356
FP	343	67
TN	13	787
FN	14	498

TABLE 6
VALUE METHOD NAÏVE BAYES

	Before	After
Accuracy	50.74%	70.32%
Precision	30.99%	64.41%
Recall	55.09%	90.86%

TABLE 7
VALUE METHOD SUPPORT VECTOR MACHINE

	Before	After
Accuracy	70.50%	66.92%
Precision	48.15%	61.31%
Recall	3.67%	92.15%

TABLE 8
VALUE ACCURACY AND AUC

	Accuracy	AUC
NB	50.74%	0.771
NB+SMOTE	70.32%	0.954
SVM	70.50%	0.508
SVM+SMOTE	66.92%	0.832

Based on Table 9 we can know that the accuracy of the Naïve Bayes Algorithm method value 50.74% shows that the accuracy obtained is included in the quite good category. And the accuracy results after using the SMOTE Up-sampling in Table 10 there is an increase to 70.50% shows that the accuracy results obtained, the NB algorithm method is very appropriate to use SMOTE optimization included in either category.

TABLE 9
ACCURACY NB

Accuracy: 50,74% +/- 2,64% (micro average: 50,74%)

	Positive (Actual)	Negative (Actual)	Class Precision
Positive (Prediction)	418	160	72,32%
Negative (Prediction)	436	196	31,01%
Class Recall	48,95%	55,06%	

Based on Table 11. We can know that the accuracy of the Support Vector Machine Algorithm method value 70.50% shows that the accuracy results obtained are included in either category. And the accuracy results after using SMOTE Up-sampling in Table 12 there is a decrease to 66.92% shows that the accuracy results obtained, the SVM algorithm method is not appropriate to use SMOTE optimization the results show are quite good.

TABLE 10
ACCURACY NB WITH SMOTE

Accuracy: 70,32% +/- 2,11% (micro average: 70,32%)

	Positive (Actual)	Negative (Actual)	Class Precision
Positive (Prediction)	425	78	84,49%
Negative (Prediction)	429	776	64,40%
Class Recall	49,77%	90,87%	

TABLE 11
ACCURACY SVM

Accuracy: 70,50% +/- 0,98% (micro average: 70,50%)

	Positive (Actual)	Negative (Actual)	Class Precision
Positive (Prediction)	840	343	71,01%
Negative (Prediction)	14	13	48,15%
Class Recall	98,36%	3,65%	

TABLE 12
ACCURACY SVM+SMOTE

accuracy: 66,92% +/- 3,04% (micro average: 66,92%)

	Positive (Actual)	Negative (Actual)	Class Precision
Positive (Prediction)	356	67	84,16%
Negative (Prediction)	498	787	61,25%
Class Recall	41,69%	92,15%	

Figure 7 illustrates a graph of the Area Under Curve (AUC) optimistic result of the validation of the Naïve Bayes Algorithm method of 0.771. We can know that the results of this AUC show the acquisition of values that fall into the average category (0.70-0.80). And the results of the validation of the Area Under Curve (AUC) graph after using the SMOTE Up-sampling in Figure 8 there was an increase to 0.954, this shows that the results obtained, the NB algorithm method is very appropriate using SMOTE optimization the results show are included in the very good category (0.90-1.00).

Figure 9 illustrates a graph of the Area Under Curve (AUC) optimistic result of the validation of the Support Vector Machine Algorithm method of 0.508. We can know that the results of this AUC show the acquisition of values that fall into the failure category (0.50-0.60). And the results of the validation of the Area Under Curve (AUC) graph after using the SMOTE Up-sampling in Figure 10 there was an increase to 0.832, this shows that the results obtained, the NB algorithm method is very precise using SMOTE optimization the results show are included in either category (0.80-0.90).

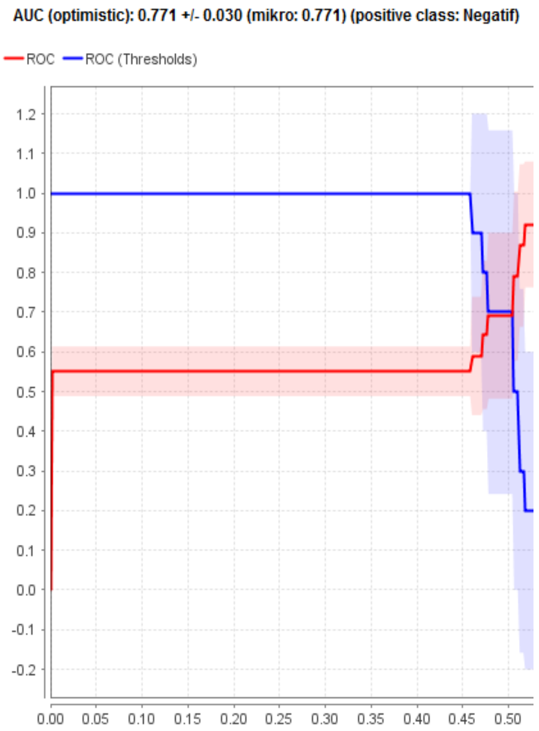

Figure 7. AUC NB

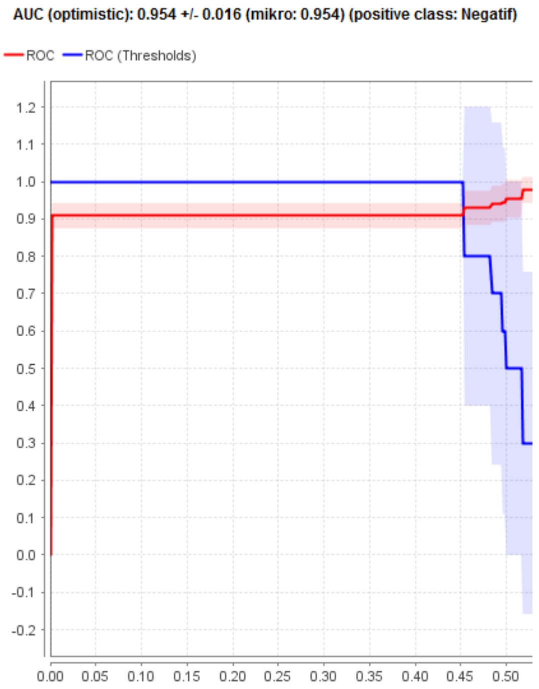

Figure 8. AUC NB with SMOTE

Comparison graph of accuracy value and AUC value between the four algorithms as presented in Figure 11 and Figure 12.

Based on the recapitulation analysis shows that the evaluation using the NB algorithm with SMOTE is the best solution to get the highest accuracy and AUC values.

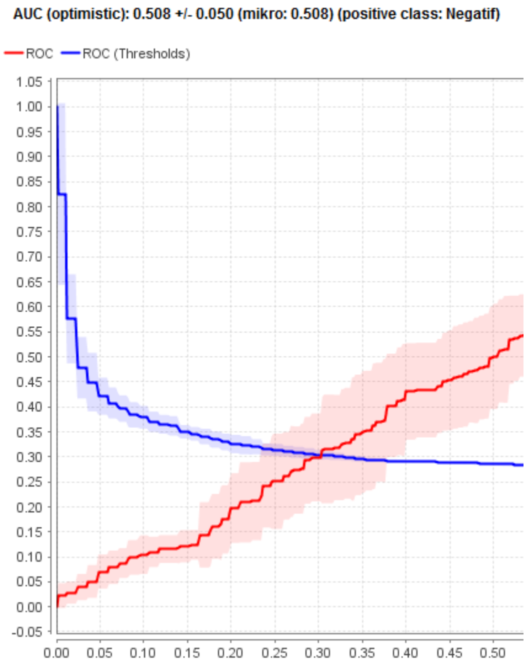

Figure 9. AUC SVM

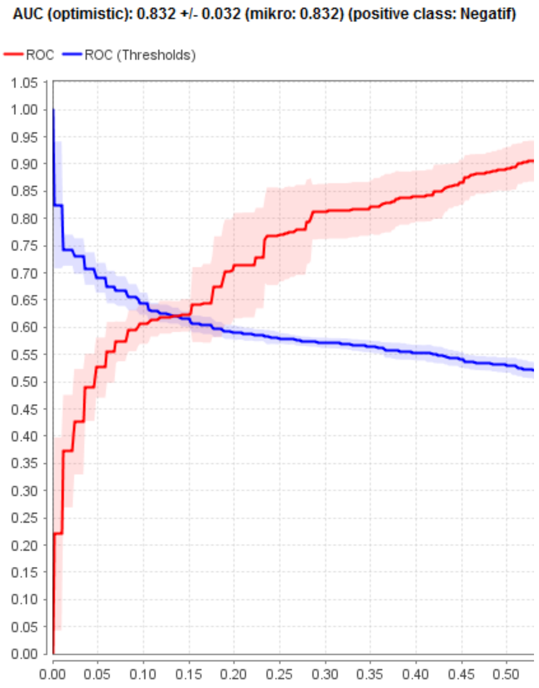

Figure 10. AUC SVM with SMOTE

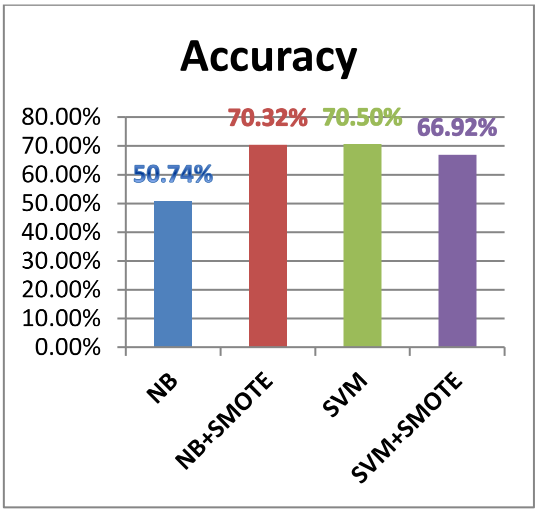

Figure 11. Accuracy Comparison Charts

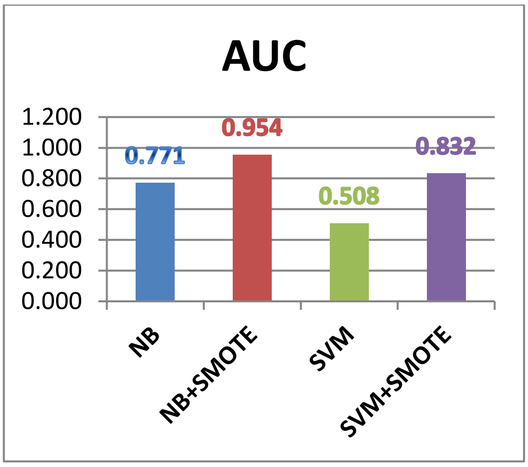

Figure 12. AUC Comparison Charts

4. Conclusion

From the results of research on the application of sensitivity classification algorithms for e-sports analysis of the training curriculum, it can be concluded that the Naïve Bayes Algorithm method, called the accuracy value of 70.32%, 64.41% precision and 90.36% recall results. Note that when the Support Vector Machine Method Algorithm is optimized using the SMOTE method, the accuracy value is 66.92%, the precision is 61.31% and the recall value is 92.15%. Based on the results of the study, it can be concluded that the Naïve Bayes algorithm has a higher accuracy than the Support Vector Machine algorithm, so that the difference in accuracy between Naïve Bayes and the Support Vector Machine can be observed at 3.4%. Thus, the Naïve Bayes algorithm can predict better when analyzing sentiment on e-sports for student learning curriculum.

As for suggestions for the continuation of this study, it is expected that in subsequent studies, more records are used so that comparisons on accuracy can be better. It is also hoped that further research can be developed with different methods or develop this research with optimization methods from compared algorithms such as Particle Swarm Optimization or other optimization methods.

References

[1] A. Waldi And I. Irwan, “Pembinaan Karakter Siswa Melalui Ekstrakurikuler Game Online E-Sports Di SMA 1 PSKD Jakarta,” *Journal Moral and Civic Education*, vol. 2(2), pp. 92–101, 2018.

[2] Gata, Basri, Baharuddin, Tohari, Hidayat, Patras Y E, Fatmasari, and Wardhani, “Algorithm Implementations Naïve Bayes, Random Forest. C4.5 On Online Gaming For Learning Achievement Predictions,” in *Proceedings of the International Conference on Research of Educational Administration and Management (ICREAM)*, 2018, doi: 10.2991/Icream-18.2019.1.

[3] C. Hadzinsky, “Industry Of Video Games. Past, Present, And Yet To Come,” Ph.D Thesis, *Scholarship @ Claremont, Claremont Colleges* P. 44, 2014, California, 2014.

[4] R. Syahran, “Ketergantungan Online Game Dan Penanganannya,” *Jurnal Psikologi Pendidikan Dan Konseling*, vol. 1(1), 2015, doi: 10.26858/Jpkk.V1i1. 1537.

[5] R. Y. Lestari, “Peran Kegiatan Ekstra Kurikuler Dalam Mengembangkan Watak Kewarganegaraan Peserta Didik,” *Untirta Civic Education Journal*, vol. 1(2), pp. 136–152, 2016, doi: 10.30870/Ucej.V1i2. 1887.

[6] E. Julius, Honggowidjaja, and P. E. Dora, “Perancangan Interior Fasilitas E-Sports Arena,” *Jurnal Intra*, vol. 4(2), pp. 672–681, 2016.

[7] S. Y. Saputra, “Permainan Tradisional Vs Permainan Modern Dalam Penanaman Nilai Karakter Di Sekolah Dasar,” *Elementary School Education Journal*, vol. 1(1), pp. 1–7, 2017.

[8] Z. Fadli, “Membentuk Karakter Anak Dengan Olahraga Tradisional,” *Jurnal Ilmu Keolahragaan*, vol. 13(2), pp. 38–44, 2014.

[9] K. Harahap And I. Beydha, “Game Online Dan Prestasi Belajar (Studi Korelasional

Pengaruh Game Online Terhadap Prestasi Belajar Siswa Kelas VIII," *Jurnal Universitas Sumatera Utara*, pp. 1–10, 2015.

[10] R. A. Amanda, "Pengaruh Game Online Terhadap Perubahan Perilaku Agresif Remaja Di Samarinda," *Jurnal Ilmu Komunikasi*, vol. 4(3), pp. 290–304, 2016.

[11] N. Husna, E. Normelani, And S. Adyatma, "Hubungan Bermain Games Dengan Motivasi Belajar Siswa Sekolah Menengah Pertama (SMP) Di Kecamatan Banjarmasin Barat," *JPG (Jurnal Pendidikan Geografi*, vol. 4(3), pp. 1–14, 2017.

[12] T. Jap, S. Tiatri, E. S. Jaya, And M. S. Suteja, "The Development Of Indonesian Online Game Addiction Questionnaire," *Plos One*, vol. 8(4), pp. 4–8, 2013, doi: 10.1371/Journal.Pone.0061098.

[13] S. Virlia And S. Setiadji, "Hubungan Kecanduan Game Online Dan Keterampilan Sosial Pada Pemain Game Dewasa Awal Di Jakarta Barat," *Jurnal Psibernetika*, vol. 9(2), 2017, doi: 10.30813/Psibernetika.V9i2.460.

[14] D. Rahmawati, D. Mulyana, S. Karlinah, Hadisiwi, "The Cultural Characteristics Of Online Players In The Internet Cafes Of Jabodetabek, Indonesia," *Journal of Theoretical and Applied Information Technology*, vol. 96(7), pp. 1868–1883, 2018.

[15] A. M. I Putu Arika Mulyasanti Pande, "Hubungan Kecanduan Game Online Dengan Prestasi Belajar Siswa Smp Negeri 1 Kuta Ni Putu Arika Mulyasanti Pande Dan Adijanti Marheni," *Jurnal Psikologi Udayana*, vol. 2(2), pp. 163–171, 2015.

[16] A. Nur, Muhammad, Hariani, Sri, Rosita, "Pengaruh Keaktifan Berorganisasi, Bimbingan Orang Tua, Kedisiplinan Belajar Terhadap Prestasi Belajar Mahasiswa Pendidikan Ekonomi Universitas Kanjuruhan Malang," *Jurnal Riset Pendidikan Ekonomi*, vol. 1(1), pp. 4–29, 2016.

[17] V. Effendy, Adiwijaya, And Baizal, "Handling Imbalanced Data In Customer Churn Prediction Using Combined Sampling And Weighted Random Forest," in *International Conference on Information and Communication Technology*, 2014, pp. 325–330, doi: 10.1109/Icoict.2014.6914086.

[18] B. Afeni, T. Aruleba, And I. Oloyede, "Hypertension Prediction System Using Naive Bayes Classifier," *Journal of Advances in Mathematics and Computer Science*, vol. 24(2), pp. 1–11, 2017, doi: 10.9734/Jamcs/2017/35610.

[19] L. Dong, X. Li, And G. Xie, "Nonlinear Methodologies For Identifying Seismic Event And Nuclear Explosion Using Random Forest, Support Vector Machine, And Naive Bayes Classification," *Hindawi publishing corporation*, vol. 2014, 2014, doi: 10.1155/2014/459137.

[20] D. Salmu, S. And A. Solichin, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta Prediction Of Timeliness Graduation Of Students Using Naïve Bayes : A Case Study At Islamic State University Syarif Hidayatullah Jakarta," in *Prosiding Seminar Nasional Multidisiplin Ilmu*, pp. 2017, 701–709.

[21] N. Nasution, K. Djahara, And A. Zamsuri, "Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes (Studi Kasus : Fasilkom Unilak)," *Jurnal Teknologi Informasi & Komunikasi Digital Zone*, vol. 1(1), pp. 1–11, 2015.

[22] S. Syarli And A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi)," *Jurnal Ilmiah Ilmu Komputer*, vol. 2(1), pp. 22–26, 2016.

[23] H. Widayu, Darma, Silalahi, And Mesran, "Data Mining Untuk Memprediksi Jenis Transaksi Nasabah Pada Koperasi Simpan Pinjam Dengan Algoritma C4.5," *Media Informatika Budidarma*, vol 1, 2017.

[24] Andriani, "Sistem Pendukung Keputusan Berbasis Decision Tree Dalam Pemberian Beasiswa Studi Kasus : Amik Bsi Yogyakarta," in *Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA 2013)*, 2013, pp. 163–168.

[25] A. Utami, "Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization," *Jurnal Pilar Nusa Mandiri*, vol. 13(1), pp. 103–112, 2017.

[26] N. Lutfiyana, "Penerapan Algoritma C4.5 Berbasis Particle Swarm Optmization Untuk Prediksi Hasil Layanan Kemudaha Donasi Zakat Dan Program," *Jurnal Pilar Nusa Mandiri*, vol. 14(1), pp. 103–110, 2018.

[27] Brown, "Data Mining For Dummies," *John Wiley & Sons, Inc., Canada* P. 381, 2013, doi: 10.1007/978-1-4614-7669-6.

[28] J. Mathew, M. Luo, C. K. Pang, And H. L.

Chan, “Kernel-Based SMOTE For SVM Classification Of Imbalanced Datasets,” in *IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society*, 2015, pp. 1127–1132, doi: 10.1109/ IECON.2015.7392251.

[29] Z. Li, X. Liu, N. Xu, And J. Du, “Experimental Realization Of A Quantum Support Vector Machine,” *Physical Review Letters*, vol. 114(14), pp. 1–5, 2015, doi: 10.1103/Physrevlett.114. 140504.

[30] S. Mehra And T. Choudhury, “Sentiment Analysis Of User Entered Text,” in *Proceedings International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018*, 2018, pp. 457–461, doi: 10.1109/Ctems. 2018.8769136.

[31] B. Pratama *Et Al.*, “Sentiment Analysis Of The Indonesian Police Mobile Brigade Corps Based On Twitter Posts Using The Svm And Nb Methods,” *Journal of Physics: Conference Series*, vol. 1201(1), 2019, doi: 10.1088/1742-6596/1201/1/ 012038.

[32] M. Fernández-Gavilanes, T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, And F. Javier González-Castaño, “Unsupervised Method for Sentiment Analysis In Online Texts” *Expert System with Applications*, vol. 58, pp. 57-75 2016.

[33] N. Colneric And J. Demsar, “Emotion Recognition On Twitter: Comparative Study And Training A Unison Model,” *IEEE Transactions Affective Computing*, vol. 3045, 2018, doi: 10.1109/Taffc.2018. 2807817.

[34] A. Giachanou And F. Crestani, “Like It Or Not: A Survey Of Twitter Sentiment Analysis Methods,” *Association for Computing Machinery*, vol. 49(2), 2016, doi: 10.1145/2938640.

[35] A. Shelar And C. Y. Huang, “Sentiment Analysis Of Twitter Data,” in *Proceedings International Conference Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 1301–1302, doi: 10.1109/Csci46756.2018.00252.

[36] P. Barnaghi, P. Ghaffari, And J. G. Breslin, “Opinion Mining And Sentiment Polarity On Twitter And Correlation Between Events And Sentiment,” in *Proceedings IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 2016, pp. 52–57, doi: 10.1109/Bigdataservice. 2016.36.

[37] Sarlan, Nadam, And Basri, “Twitter Sentiment Analysis,” in *International Conference on Information Technology and Multimedia*, 2014, pp. 212–216, doi: 10.1109/Icimu.2014.7066632.

[38] Z. Jianqiang And G. Xiaolin, “Comparison Research On Text Pre-Processing Methods On Twitter Sentiment Analysis,” *IEEE Access*, Vol. 5, Pp. 2870–2879, 2017,
Doi: 10.1109/Access.2017.2672677.

[39] Effrosynidi, Symeonidis, And Arampatzis, “A Comparison Of Pre-Processing Techniques For Twitter Sentiment Analysis,” in *International Conference On Theory and Practice of Digital Libraries*, 2017, pp. 394–406, doi: 10.1007/978-3-319-67008-9.

[40] Saif, Y. He, M. Fernandez, And H. Alani, “Contextual Semantics For Sentiment Analysis Of Twitter,” *Information Processing & Management*, vol. 52(1), pp. 5–19, 2016, doi: 10.1016/J.Ipm. 2015.01.005.

[41] G. A. Buntoro, “Analisis Sentimen Calon Gubernur Dki Jakarta 2017 Di Twitter”, *Journal of Information Technology*, vol. 2(1), pp. 32–41, 2017.

[42] J. Li, S. Fong, Y. Zhuang, And R. Khoury, “Hierarchical Classification In Text Mining For Sentiment Analysis Of Online News”, in *International Conference on Soft Computing and Machine Intelligence*, vol. 20(9), pp. 3411–3420, 2016, doi: 10.1007/S00500-015-1812-4.

[43] Kamyab, Tao, Mohammadi, And Rasool, “Sentiment Analysis On Twitter: A Text Mining Approach To The Afghanistan Status Reviews,” in *Proceedings of the International Conference on Artificial Intelligence and Virtual Reality*, 2018, pp. 14–19, doi: 10.1145/3293663.3293687.