IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.14, No.4, October 2020, pp. 429~440
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.58579

Twitter's User Opinion About Master and Doctoral Degrees: A Model of Sentiment Comparison

Victor Wiley*[1], Thomas Lucas[2]
[1,2]Cemerlang research, Indonesia
e-mail: ***[1]victorwiley10@gmail.com**, [2]thomasreliable10@gmail.com

Abstrak

Makalah ini mengkaji pendapat calon mahasiswa tentang rencana mereka melanjutkan studi ke jenjang magister (S2) dan doktor (S3). Kurangnya pendekatan dalam mencari opini publik tentang minat calon mahasiswa untuk melanjutkan studi ke jenjang yang lebih tinggi seperti magister atau doktoral. Melalui tulisan ini, opini pengguna Twitter diekstraksi menggunakan teknik data mining tertentu untuk mengetahui tiga jenis sentimen (negatif, netral, dan positif) dengan mengambil jenis emosi yang paling dominan (yaitu kemarahan, antisipasi, cinta, ketakutan, kegembiraan, kesedihan, kejutan, kepercayaan). Dataset dibagi menjadi dua kelompok pengguna Twitter. Kedua dataset tersebut mewakili kelompok A. Pendapat tersebut tentang melanjutkan studi ke jenjang magister versus kelompok B yang melanjutkan ke jenjang doktor. Kelompok tersebut kemudian dibagi menjadi tiga jenis pernyataan sentimen tentang gelar master versus gelar doktor. Kelompok pertama adalah sentimen melanjutkan studi ke jenjang magister dengan hasil: (a) 109 tweet negatif, 1683 tweet netral dan 131 tweet positif. Untuk kelompok kedua (mis., Sentimen mahasiswa tentang melanjutkan ke gelar doktor), memiliki hasil: (a) 421 tweet negatif, 7666 tweet netral dan 1805 tweet positif. Data yang diujikan memberikan nilai akurasi 85%. Hasil analisis sentimen ini berguna sebagai referensi bagi perguruan tinggi untuk memahami perkembangan sentimen (opini) dari pengguna Twitter dan membantu institusi untuk meningkatkan reputasi dan kualitasnya.

Kata kunci*— sentimen, Naïve Bayes, twitter*

Abstract

This paper examines the opinion of student candidate about their plan to study further to master degree (S2) and doctoral degree (S3). There is lack of approach in finding public opinion about the interest of student candidate in continuing study to higher level such as master degree or doctoral degree. Through this paper, the Twitter's user opinions are extracted using certain data mining technique to find out three sentiment types (negative, neutral, and positive) by taking the most dominant type of emotions (i.e., anger, anticipation, love, fear, joy, sadness, surprise, trust). The dataset is divided into two groups of Twitter's users. Both datasets represent group A those opinion is about continuing study further to master degree versus group B whose continuing to doctoral degree. The groups are then divided into three types of sentiment statements about master degree versus doctoral degree. The first group is their sentiment about continuing study further to master degree with the result: (a) 109 negative tweets, 1683 neutral tweets and 131 positive tweets. For the second group (e.g., student's sentiments about continuing to doctoral degree), it has results: (a) 421 negative tweets, 7666 neutral tweets and 1805 positive tweets. The data are tested to give accuracy value of 85%. The result of this sentiment analysis is useful as a reference for universities to understand the development of sentiments (opinion) from Twitter's users and help the institutions to improve their reputation and quality.

Keywords— *sentiment, Naïve Bayes, twitter*

Received August 13th,2020; Revised October 27th, 2020; Accepted October 30th, 2020

1. INTRODUCTION

Higher education institutions experienced a higher pressure to maintain their quality and reputation. They have to shift their educational paradigm from an authoritarian to consumer-friendly perspectives. There is a large lag experienced by many universities that they are powerful in knowledge but lacking students' enrollment. This vacancy occurs in many master degrees (S2) and doctoral degrees (S3). Such tertiary institutions are characterized by having large capacities, however, some of them only gain a small number of students enrollment.

Although they have implement brilliant strategies, however, with the fast development of smart and young millennial population, such tertiary education institutions sometimes is not updated to understand the sentiment (opinion) of the young community.

As the young millennial population grows, the demand of education quality becomes more challenged especially among the tertiary institutions. In fact, previous studies reported that some tertiary institutions often do not understand why people do not choose and are interested in continuing their studies at their institutions. Although there are efforts in improving their education quality, the universities lack information about the right way to start gaining new students into their campus.

There is rare research and information in explaining this issue especially in the recent scholarly body of knowledge.

Some studies showed that many higher education institutions lack of supports from both students and the community. In addition, some scholars[1] showed that universities needs adequate and reliable information to understand the student candidates in order they can gain and drive the student to be willing to enter and enroll to their university.

In addition, obtaining information directly from graduates S1 or S2 is often not easy[2]. Most of them will reluctant to explain their future education plan. They may not be willing to open debatable opinions directly to universities for various reasons. Although there have been many surveys of interest in continuing study further, such studies often have limited demographic scope so that they do not represent the general views of graduates and prospective students.

Thus, there is an information vacuum to understand the views and experiences of students and alumni related to the quality of tertiary education institution and what steps that universities must take to maintain the quality of their education.

One source of information that is considered rich-content and always being updated is Twitter social media. Twitter has produced 110 million tweets every day and has more than 200 million users [3]. As big data source used by globally diverse population, most Twitter's users are teenagers and young adults. The users interact actively and talk about many topics. They collaborate to create join contents and spread messages (tweets) to their network using local slang language and culture. They can talk about their personalities and experiences and expand information to their preferred social network.

Given the variety of local cultures, it is often difficult to do data mining on tweets that have various forms of language and dialect[4]. To support the purpose of this research, namely in the Indonesian context, the tweets will be taken in Indonesian.

1.1. Problem formulation

Tweets from Twitter have various forms of meaning and sentiment[3]. There is a main difficulty to pull information and use the tweets to extract for information. To facilitate extraction, it needs dataset cleaning and classification so that the large data such as tweets can be simplified through appropriate classification methods.

Machine algorithms very often help to classify and predict whether a document represents positive or negative sentiment[4]. Machine learning is categorized under two types known as supervised and unsupervised machine learning algorithms[5]. The supervised algorithm uses labeled data sets where each training set document is labeled with the appropriate sentiment[6]. Meanwhile, unsupervised learning includes unlabeled data sets in which text is not labeled with the appropriate sentiment. This study mainly discusses supervised learning techniques on labeled data sets.

Many previous studies have used this method to classify sentiments. Sarkar, et.al.,[7] conducted a comparison of the three methods of Naive Bayes, Decision Tree, and Neural Networks. Overall research results indicate that Naive Bayes is the best choice for domain training. Routray et.al.,[8] and Khairnar & Kinikar [9] discuss the many approaches of different researchers, and suggest that machine learning methods are an efficient way of analyzing sentiment.

Saif, Hassan; He, Yulan and Alani, Harith [10] developed a method that can collect corpus automatically and train a naïve bayes multinomial-based sentiment classifier using 2 features, namely n-gram and POS-tag and the classifier determines the class of each tweet.

Dey, Lopamudra & Chakraborty [11] collected 2 sets of datasets, namely movie reviews and hotel reviews using 2 classifiers of naïve bayes and K-NN. The aim was to check which classifier gave the better result on both data sets. The experimental results show that the Naïve Bayes classifier performs better in terms of the movie review data set and in considering the hotel review data set, both classifiers show approximate results. Lastly, the naïve Bayes classifier is better for the movie review classification.

There are many models and approaches for classifying data to produce information for stakeholders. One common approach is to apply the Naive Bayes algorithm [12]. The difference between the method in this study and the previous research is In this paper for Sentiment Analysis we are using two Supervised Machine Learning algorithms with Naïve Bayes', Stop word, and normalization. This algorithm is simple enough to recognize data patterns and do classifications with a good degree of accuracy.

1.2. Research purposes

The purpose of this study is to classify sentiments in tweet data and use text mining technique with the Naïve Bayes algorithm to recognize the data patterns and conducting sentiment analysis about the Twitter's user opinions. It also seeks explanation about the accuracy and performance of Naïve Bayes algorithm to extract and classify the sentiment analysis from Twitter's user's opinion.

This paper contains five parts which will be explained and presented below. Part one contains problem formulation and research purposes. Part two contains theoretical review about the Naïve Bayes algorithm, Bayes' Theorem and part three contains research methods and data mining technique. The discussion with analysis result is given in part four. Finally, the conclusion, suggestion and implication are given in the final part.

2. METHODS

This study applied data mining technique through certain problem analysis and literature study. Firstly, we have determined the formulation of the problem from the field observation. In this case, the problem is observed from the perspective of the cause of the students reluctant and how they perceive about the master and doctoral degrees. After formulating the existing problems, the scope of the problem is analyzed to find out how to solve these problems and determine the solution. We are then moved to the theoretical basis of various literatures regarding the application of the Naïve Bayes method, concepts and theories of data mining and the use of the algorithm to classify the dataset [13]. It is conducted through journal review in order to get the knowledge base and select the appropriate data mining. We divide the population into two groups of Twitter's datasets. The three types of sentiment statements about continuing study is tested for both group A which representing the student candidate of master degree versus group B of doctoral degree.

Naive Bayes is a simple probabilistic classification that calculates a set of discrete values by adding up the frequency and combination of values from a given dataset[14]. The algorithm uses the Bayes theorem and assumes all the independent or non-interdependent attributes given by values to class variables [15]. Another definition says that Naive Bayes is combined classification with probability and statistical methods presented by the British scientist Thomas Bayes, which predicts future opportunities based on experience in the past [16].

Naive Bayes is based on the simplification assumption that attribute values are conditionally mutually independent with certain given output values [17]. To achieve output value, the probability of observation is a product of individual probabilities. Naive Bayes has advantage since it uses a small number of training data to determine the estimated parameters needed in the classification process[18]. Naive Bayes often works far better in most complex real-world situations as expected [19].

To explain the Naive Bayes method, it is important to know that the classification process requires a number of clues to determine what class is suitable for the analyzed sample[20]. Therefore, the Naive Bayes method above is adjusted as follows:

$$P(C|XF1 \ldots Fn) = \frac{P(C)P(F1 \ldots Fn|C)}{P(F1 \ldots Fn)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

2.2. *Sensitivity, Specificity And Accuracy*

Three testing approach are used to test the proposed Naïve Bayes model, eg, sensitivity, specificity and accuracy[21]. They will be True positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). If a sentiment is proven to exist in a Twitter's message, the test given will indicate the presence of the opinion, and results of the test are considered to be True Positive (TP)[22]. Likewise, if a sentiment is proven to be null or void in the message, the tests indicate that the opinion is also null or void which provide the test result value as True Negative (TN). Both true positives and true negatives show consistent results between the tests and proven conditions (also called truth standards). However, there is no perfect classification to get entire Twitter's users opinion. Therefore certain criteria are implemented to get adequate result. The test shows that an opinion from a user who does not actually have such a sentiment, the test result is false positive (FP). Likewise, if the test result indicates that the sentiment does not exist for a user with a definite opinion, the test result is False Negative (FN). Both False Positive and False Negative indicate that the test results can be contrary to the actual conditions.

2.3. *Crawling for collecting data*

The systematic procedure used to collect data is by crowding out data taken from Twitter API. The crawling process has a goal to extract the data from the Twitter Big data. After the data is collected, data analysis is performed to adjust the data process to be processed in the Naïve Bayes method.

2.3.1. *Determination of keywords*

The keyword is selected based on the sentiment classification which done by breaking down emotions into 10 different emotions, e.g., anger, anticipation, love, fear, joy, sadness, surprise, belief, negative and positive emotion.

2.3.2. *Filtering for simplifying the Twitter hashtag*

A hashtag is a short special character or hash mark with a symbol '#' which placed before the word. Hashtag are used to mark certain topics that are considered important[23]. The hash tag is used to divide the message (tweets) into raw a 'text' field contains the tweet, hashtag, and URL sections. The hashtag and URL are removed or filtered to get the main text field of the tweet section[24]. It contains many URLs, hashtags, and other Twitter handles. We will delete all of this using the $gsub$ function.

2.3.3. *Initializing Stop word*

The next problem in the data mining process is so-called a Stop word. These are frequent words but provide little information. Stop word is unnecessary word which should be deleted to simplify the body of text. In some common English vocabulary, stop words include "I", "he", "people", etc. In the tm package, there are 174 Stop words in this public list. In fact, stop word must be used carefully in order not to break normal tweets due to it can cause an over-emphasized frequency analysis and lead to incorrect result interpretation.

2.3.4. Removing contextual stop words

Removing stops word is the main step in the natural language processing. The contextual stop words include 'a', 'the', 'is'. The removal is important since it will define a function and apply it to our*DataFrame*. We create a set of words that we will call ‘stops’ by using a function with of Naïve Bayes to speed up removing the stop words.

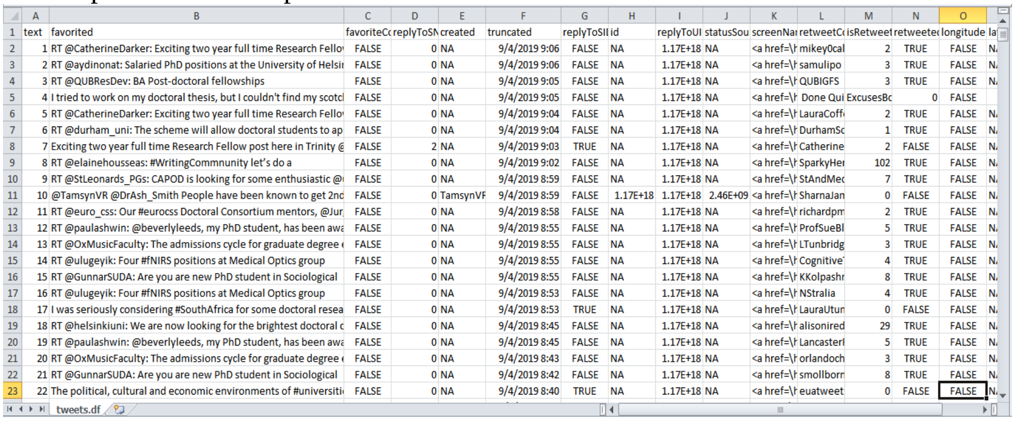
2.4 Determining software Tools for Data Analysis

The processing of the data is conducted by using a tool named RStudio, in this study the data to be processed is text dataset containing tweets to get the opinion of the Twitter’ users about their interests of continuing study further to master (S2) and doctoral (S3) degrees. The data is extracted and crawled in August 2019. The extracted data is saved in an Excel spreadsheet with a filename of *college.csv*. The file is then included and loaded into RStudio application for further processing.

3. RESULTS AND DISCUSSION

In this study, the tweets dataset have been extracted to determine the sentiment of the Twitter’s users about their intention to study further to master and doctoral degrees. The collected tweets contain their opinions and interests about the study preferences. There are 1243 tweets related to master degrees and 17831 tweets about doctoral degree. The data which needed in this study consisted of two types, e.g., training data and test data. The training dataset is categorized through the Naïve Bayes and provide final result of the negative, neutral and positive sentiments.

This study found a total of 10 variables after implementing the *userTimeline* function to give us a snapshot of the sample data as shown below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	text	favorited	favoriteCo	replyToSN	created	truncated	replyToSII	id	replyToUI	statusSou	screenNar	retweetCo	isRetweet	retweeted	longitude	la
2	1	RT @CatherineDarker: Exciting two year full time Research Fello	FALSE	0	NA	9/4/2019 9:06	FALSE	NA	1.17E+18	NA	<a href=\h	mikey0cal	2	TRUE	FALSE	N
3	2	RT @aydinonat: Salaried PhD positions at the University of Helsi	FALSE	0	NA	9/4/2019 9:06	FALSE	NA	1.17E+18	NA	<a href=\h	samulipo	3	TRUE	FALSE	N
4	3	RT @QUBResDev: BA Post-doctoral fellowships	FALSE	0	NA	9/4/2019 9:05	FALSE	NA	1.17E+18	NA	<a href=\h	QUBIGFS	3	TRUE	FALSE	N
5	4	I tried to work on my doctoral thesis, but I couldn't find my scotc	FALSE	0	NA	9/4/2019 9:05	FALSE	NA	1.17E+18	NA	<a href=\h	Done Qui	ExcusesBc	0	FALSE	
6	5	RT @CatherineDarker: Exciting two year full time Research Fello	FALSE	0	NA	9/4/2019 9:04	FALSE	NA	1.17E+18	NA	<a href=\h	LauraCoff	2	TRUE	FALSE	N
7	6	RT @durham_uni: The scheme will allow doctoral students to ap	FALSE	0	NA	9/4/2019 9:04	FALSE	NA	1.17E+18	NA	<a href=\h	DurhamSc	1	TRUE	FALSE	N
8	7	Exciting two year full time Research Fellow post here in Trinity @	FALSE	2	NA	9/4/2019 9:03	TRUE	NA	1.17E+18	NA	<a href=\h	Catherine	2	FALSE	FALSE	N
9	8	RT @elainehousseas: #WritingCommnunity let’s do a	FALSE	0	NA	9/4/2019 9:02	FALSE	NA	1.17E+18	NA	<a href=\h	SparkyHer	102	TRUE	FALSE	N
10	9	RT @StLeonards_PGs: CAPOD is looking for some enthusiastic @	FALSE	0	NA	9/4/2019 8:59	FALSE	NA	1.17E+18	NA	<a href=\h	StAndMec	7	TRUE	FALSE	N
11	10	@TamsynVR @DrAsh_Smith People have been known to get 2nd	FALSE	0	TamsynVF	9/4/2019 8:59	FALSE	1.17E+18	1.17E+18	2.46E+09	<a href=\h	SharnaJan	0	FALSE	FALSE	N
12	11	RT @euro_css: Our #eurocss Doctoral Consortium mentors, @Jur	FALSE	0	NA	9/4/2019 8:58	FALSE	NA	1.17E+18	NA	<a href=\h	richardpm	2	TRUE	FALSE	N
13	12	RT @paulashwin: @beverlyleeds, my PhD student, has been awa	FALSE	0	NA	9/4/2019 8:55	FALSE	NA	1.17E+18	NA	<a href=\h	ProfSueBl	5	TRUE	FALSE	N
14	13	RT @OxMusicFaculty: The admissions cycle for graduate degree	FALSE	0	NA	9/4/2019 8:55	FALSE	NA	1.17E+18	NA	<a href=\h	LTunbridg	3	TRUE	FALSE	N
15	14	RT @ulugeyik: Four #fNIRS positions at Medical Optics group	FALSE	0	NA	9/4/2019 8:55	FALSE	NA	1.17E+18	NA	<a href=\h	Cognitive	4	TRUE	FALSE	N
16	15	RT @GunnarSUDA: Are you are new PhD student in Sociological	FALSE	0	NA	9/4/2019 8:55	FALSE	NA	1.17E+18	NA	<a href=\h	KKolpashr	8	TRUE	FALSE	N
17	16	RT @ulugeyik: Four #fNIRS positions at Medical Optics group	FALSE	0	NA	9/4/2019 8:53	FALSE	NA	1.17E+18	NA	<a href=\h	NStralia	4	TRUE	FALSE	N
18	17	I was seriously considering #SouthAfrica for some doctoral resea	FALSE	0	NA	9/4/2019 8:53	TRUE	NA	1.17E+18	NA	<a href=\h	LauraUtun	0	FALSE	FALSE	N
19	18	RT @helsinkiuni: We are now looking for the brightest doctoral c	FALSE	0	NA	9/4/2019 8:45	FALSE	NA	1.17E+18	NA	<a href=\h	alisonired	29	TRUE	FALSE	N
20	19	RT @paulashwin: @beverlyleeds, my PhD student, has been awa	FALSE	0	NA	9/4/2019 8:45	FALSE	NA	1.17E+18	NA	<a href=\h	Lancaster	5	TRUE	FALSE	N
21	20	RT @OxMusicFaculty: The admissions cycle for graduate degree	FALSE	0	NA	9/4/2019 8:43	FALSE	NA	1.17E+18	NA	<a href=\h	orlandoch	3	TRUE	FALSE	N
22	21	RT @GunnarSUDA: Are you are new PhD student in Sociological	FALSE	0	NA	9/4/2019 8:42	FALSE	NA	1.17E+18	NA	<a href=\h	smollborn	8	TRUE	FALSE	N
23	22	The political, cultural and economic environments of #universiti	FALSE	0	NA	9/4/2019 8:40	TRUE	NA	1.17E+18	NA	<a href=\h	euatweet	0	FALSE	FALSE	N

Figure 1 Total of 10 variables after implementing *userTimeline* function
Source: Analysis result with RStudio

The ‘text’ field contains the tweet, hashtag, and URL sections. We need to remove the hashtag and URL from the text field so we only have the main tweet section to run our sentiment analysis. Our current text fields look like below:

3.1. Sentiment Analysis Model

The probability value of each criterion is obtained from the output of big data extraction results in table 1. The probability value of each criterion is as follows. Based on table 1 and table 2 data, the data is then filtered with a *get_sentiment* function to extract sentiment scores for each tweet. We obtained output that shows a variety of negative emotions that exist in each tweet (Table 2).

3.2. Negative Sentiment

The negative statement about continuing study is tested for both group A which representing the student candidate of master degree versus group B of doctoral degree.

Table 1 Tweets with negative sentiments about continuing study to master degrees

No	Tweet statement
1	"No offense but lulus kuliah tahun ini actually sucks! https://t.co/54Cfxbyhzy"
2	nUpcoming calendar for kuliah on September 2019...\nMark your calendar as you won't want to miss t… https://t.co/vlnNnkD2Kn"
3	"imagine officially being a junior highschool teacher lmao but my soul still anak kuliah tongkrongan. https://t.co/7QD3EolMC3"
4	Kuliah lah nder, it will be your best way to get a best job then you can buy your iphone with your own money"
5	I expected all dewan kuliah are like DK D1 DK D2 and all DKss but
6	no offense but ganti jadwal kuliah is actually sucks"

Source: Analysis result with RStudio

Table 2 Tweets with negative sentiments about continuing study to doctoral degrees

No	Tweet statement
1	Doctoral students disproportionately experience anxiety, depression, and other forms of mental illness throughout… "
2	"Run.... Never back !!! You are fucking COWARD !!\n "
3	I tried to work on my doctoral thesis, but I couldn't find my scotch tape.
4	BA Post-doctoral fellowships\nif you are interested in applying to come and work at Queen's get in touch https://t.co/B0vYi7T…

Source: Analysis result with RStudio

As showed in Table 2, there are some opinions with the meaning of negative sentiment. For example, on line one there are words that mean that the Twitter user feels bored about their study or continuing further to future study. Then, on line two there is negative sentiment which means that the user feels bored about their research progression. Also, in line 2 there is an opinion that the no.2 user feels desperation about graduating from college due to short time of study.

3.2.1. Neutral Sentiment

Neutral statement about continuing study is tested for both group A which representing the student candidate of master degree versus group B of doctoral degree. In the neutral dataset tweet group, we display some examples of text representing student candidate of master degree and doctoral degree which can be seen in Table 3 and Table 4.

Table 3 Tweets with neutral sentiment about continuing study to master degrees

No	Tweet statement
1	RT
2	Untung ntar mo kuliah onlen. gaperlu dah tuh ketemu org2 yg permasalahin outfit cuz u wont see me bitch. i be lying…
3	kenapa mau jualan? Ga ribet berbisnis sambil kuliah? \n<U+0001F469> : bcs i enjoyed it, i like to share something that some…
4	nsenin masuk kuliah auto makan nasi + garem
5	whenever ada kuliah tamu, i feel stupid lol

Source: Analysis result with RStudio

Table 4 Tweets with neutral sentiment about continuing study to doctoral degrees

No	Tweet statement
1	Exciting two year full time Research Fellow post here in Trinity @tcddublin. Funded by @hse Mental Health services. Wo…"
2	Salaried PhD positions at the University of Helsinki (@helsinkiuni). https://t.co/cCCUPhCgon"
3	BA Post-doctoral fellowships\nif you are interested in applying to come and work at Queen's get in touch https://t.co/B0vYi7T…"
4	Exciting two year full time Research Fellow post here in Trinity @tcddublin. Funded by @hse Mental Health services. Wo…"
5	The scheme will allow doctoral students to apply for a 3-month placement at POST to gain valuable policy experience. https:…"

Source: Analysis result with RStudio

3.2.2. Positive Sentiment

The positive statement about continuing study is tested for both group A which representing the student candidate of master degree versus group B of doctoral degree. Grouping of tweet message is based on positive tweets among Tweeter's user who comment about continuing studying further to master degree (S2). Their positive sentiment is extracted and simplified as in Table 4.

Table 5 Tweets with positive sentiment about continuing study to master degrees

No	Tweet statement
1	kenapa mau jualan? Ga ribet berbisnis sambil kuliah? \n\U0001f469 : bcs i enjoyed it, i like to share something that some… "
2	nsenin masuk kuliah auto makan nasi + garem"
3	"gue lulus kuliah mau balik indo and make a 24 hour boba place not because i think it'll make me a lot of money but bc I Want That"
4	"At certain point, I should agree that kuliah is a comfort zone."
5	"i know this won't change anything but i just hope they come to indo on feb 2020 so i can watch it klo sept tuh kuli… "
6	"Kuliah di Seoul Absence of\nComprehensive Art School."

Source: Analysis result with RStudio

Table 6 Tweets with positive sentiment about continuing study to doctoral degrees

No	Tweet statement
1	"What a great morning in Dubrovnik! ☀□ Shiny sun &amp; bright ideas about better understanding of wellbeing in Doctoral…
2	"Job offer for Doctoral Thesis "
3	"KBS World published an interview with me. It covered my doctoral research on the Bowiseong. I am grateful to KBS fo… "
4	"RT billfreehomes \"RT durham_uni \"The scheme will allow doctoral students to apply for a 3-month placement at POST t…
5	"i know this won't change anything but i just hope they come to indo on feb 2020 so i can watch it klo sept tuh kuli… "

Source: Analysis result with RStudio

Next we calculate the score of each tweet. In total, 1923 tweets were evaluated, so there must be 1923 positive, negative or neutral scores, one for each tweet. So, now it can be seen that the total score of tweets for the interest of S1 graduates to continue their studies at S2 and the interest of S2 graduates to continue to S3 is obtained 17831 total tweets.

3.3. Scores for three groups of sentiment tweets

The probability criteria for the number of dependents are converted into three numerical number, e.g., a negative sentiment then it is negative as 0, neutral is 1, or positive is 2. We then get an emotional score for each tweet as frequency number set as show in the table below. The tweets are grouped based on the 10 different emotions, e.g., anger, anticipation, love, fear, joy, sadness, surprise, trust, negative and positive.

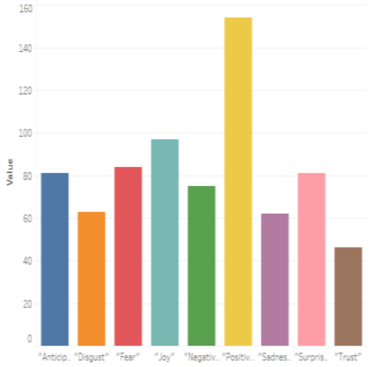


Figure 2 Emotional Score

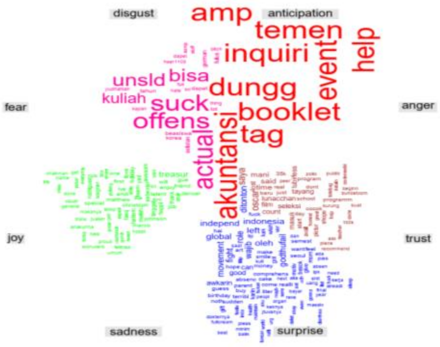


Figure 3 Emotional classification based on tweet results

Source: Analysis result with RStudio

By breaking down emotions into 10 different emotions (e.g., anger, anticipation, love, fear, Joy, sadness, surprise, trust, negative and positive), we get ten groups of emotion which will be converted to the numerical data. The process of numericalization is explained in the next step.

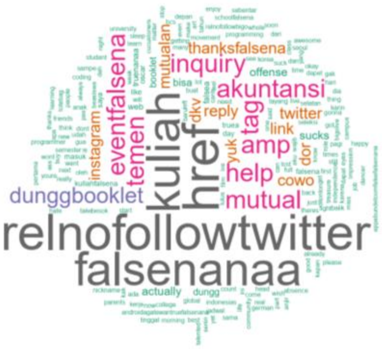


Figure 4 Wordcloud tweets for the interest of S1 graduates to continue their studies to master degree (S2)

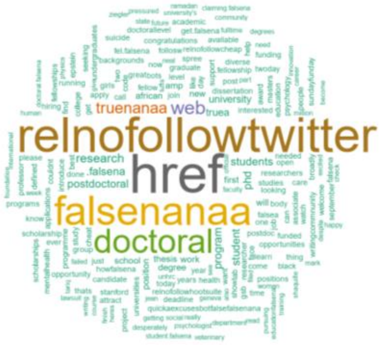


Figure 5 Wordcloud tweets for the interest of S2 graduates to continue their studies to master degree (S3)

Source: Analysis result with RStudio

We finally get cloudword of three sentiment class (e.g., positive, neutral, negative) for both group A which representing the student candidate of master degree versus group B of doctoral degree.

3.4. Numericalization

The results of numericalization is using Bayes theorem to generate negative numerical data which represents that the sentiment is a negative sentiment. Furthermore, we convert the number data into string with intParseToString coding in RStudio to get the data which presented in Table 7.

Table 7 Grouping based on classification tweets with Bayes

[1] -1.00 0.40 0.25 0.00 0.50 -1.00 0.00 0.00 0.00 0.00 0.00 0.00 -0.75
[14] 0.00 0.40 0.40 0.30 0.80 0.80 0.80 0.00 -0.35 0.00 -0.15 -1.45 0.60
[27] -0.50 0.50 -1.00 0.00 0.00 1.40 0.00 0.50 0.00 -0.50 0.00 0.25 0.00
[40] 1.30 0.00 0.00 -1.80 0.00 0.00 0.00 0.85 0.00 0.00 1.55 0.25 0.25
[53] 0.00 0.00 0.00 0.80 0.75 -1.00 0.00 0.00 0.00 0.00 -0.25 0.00 0.00
[66] 0.00 -0.25 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
[79] 0.00 0.00 0.00 0.00 0.00 0.00 -0.20 0.00 -0.80 0.80 1.05 0.00 0.00
[92] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 -0.60 0.00 0.00 0.00 0.00 0.00

Source: Analysis result with RStudio

Table 8 The list of classification result of negative, positive and neutral sentiment

[1] "Neutral" "Negative" "Positive" "Neutral" "Positive" "Negative" "Negative" "Neutral"
[9] "Neutral" "Neutral" "Neutral" "Neutral" "Neutral" "Neutral" "Neutral" "Neutral"
[17] "Neutral" "Positive" "Neutral" "Positive" "Neutral" "Positive" "Neutral" "Neutral"
[25] "Neutral" "Neutral" "Neutral" "Neutral" "Positive" "Neutral" "Positive" "Positive"
[33] "Neutral" "Neutral" "Neutral" "Neutral" "Neutral" "Neutral" "Positive" "Neutral"
[41] "Neutral" "Positive" "Neutral" "Negative" "Neutral" "Neutral" "Neutral" "Negative"
[49] "Neutral" "Negative" "Neutral" "Positive" "Neutral" "Neutral" "Positive" "Negative"

Source: Analysis result with RStudio

Finally, we tested the three groups of sentiments about continuing study to master degrees toward the doctoral degrees. The result is presented in Figure 6 and Figure 7.

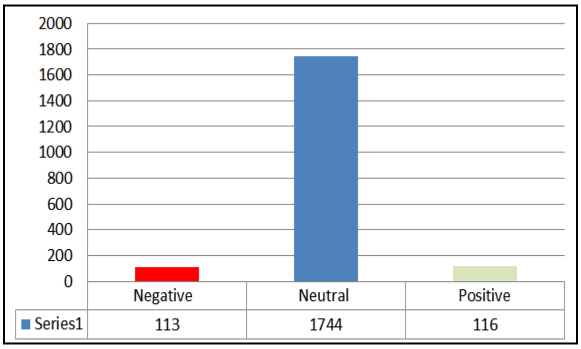


Figure 6 Tweets with three groups of sentiments about continuing study to master degrees

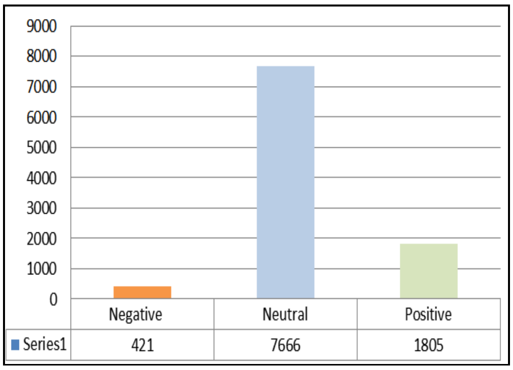


Figure7 Tweets with three groups of sentiments about continuing study to doctoral degrees

Source: Analysis result with RStudio

3.5. Testing the accuracy of the proposed model

After classification and Naïve Bayes implementation, we want to know that the proposed model with provide adequate accuracy result. The accuracy phase is useful to know the performance of the proposed model after it is implemented the Naïve Bayes classifier algorithm in classifying text data. The accuracy testing will provide information about the model performance either it will give high accuracy or low accuracy. The following table (Table 9) shows the testing result of the accuracy of the proposed models. The first model represents group A of student candidate of master degree versus group B of doctoral degree.

Table 9 Testing result of the accuracy of the proposed models

Model for Group A	Model for Group B
Accuracy : 0.78	Accuracy : 0.80
95% CI : (0.7728, 0.8435)	95% CI : (0.7863, 0.6765)
No Information Rate : 0.53	No Information Rate : 0.75
P-Value [Acc > NIR] : <2e-16	P-Value [Acc > NIR] : <2e-34
Kappa : 0.4474	Kappa : 0.5744
Mcnemar's Test P-Value : 0.1383	Mcnemar's Test P-Value : 0.2463
Sensitivity : 0.7453	Sensitivity : 0.7652
Specificity : 0.7702	Specificity : 0.7847
Pos Pred Value : 0.8038	Pos Pred Value : 0.8437
Neg Pred Value : 0.8153	Neg Pred Value : 0.8436
Prevalence : 0.4300	Prevalence : 0.4531
Detection Rate : 0.4540	Detection Rate : 0.4536
Detection Prevalence : 0.5460	Detection Prevalence : 0.5537
Balanced Accuracy : 0.7677	Balanced Accuracy : 0.7954

Source: Analysis result with RStudio

We have tested two groups of Twitter's datasets. The three types of sentiment statements about continuing study is tested for both group A which representing the student candidate of master degree versus group B of doctoral degree. The first group represents the opinion from the Twitter's user who sent their tweets containing sentiments of study further to master degree (S2). For the second group, it represents the opinion from the Twitter's user who sent their tweets containing sentiments of study further to doctoral degree (S3). From Table 9, it found that the accuracy test results can predict about the first group with 78% compared to the second group.

4. CONCLUSIONS

Based on the results of the analysis and testing that has been done, it can be concluded that the proposed model which containing the Naïve Bayes algorithm is quite successful in predicting the correct sentiment category. The testing result of both groups using the Twitter's datasets showed that the three types of sentiment statements about continuing study is important for both group A and group B. It is evidenced from the accuracy of the naïve Bayes algorithm has provided a high accuracy of 78% and 80%, which means that the model can classify text data very well. This means that the model can be used to classify tweets and also predict the user sentiment.

The author provides suggestion about the result of the study. In order to increase the quality and performance of the model, it is important to increase the amount of training data and test data to get better results when classifying tweets. In addition, the scope of the classification can be expanded to other educational institutions. Finally, it is suggested for future studies can combine the Naïve Bayes algorithm with other classification algorithm such as support vector machine and other text classification algorithms.

The proposed model is prospective to be used for practical implementation for tertiary higher education institution in monitoring and developing strategies to increase their student enrollment. In addition, the study result can bring benefit for the institutions to maintain further interest in college and foster the interest of study among prospective student. The higher education can build interest and motivation, especially strategies to interact with the community so that the reputation and brand of the university can be more popular with a good reputation.

REFERENCES

[1] Zhao, Jichang, Li Dong, Junjie Wu, and Ke Xu. "Moodlens: an emoticon-based sentiment analysis system for chinese tweets." In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1528-1531, 2012. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2339530.2339772 [Accessed: 17-Des-2019]

[2] Farzindar, Atefeh, and Diana Inkpen. "Natural language processing for social media." Synthesis Lectures on Human Language Technologies 8, no. 2 ,2015 [Online]. Available: https://www.morganclaypool.com/doi/abs/10.2200/S00659ED1V01Y201508HLT030 [Accessed: 17-Des-2019]

[3] Shoukry, Amira, and Ahmed Rafea. "Preprocessing Egyptian dialect tweets for sentiment mining." In *The Fourth Workshop on Computational Approaches to Arabic Script-based Languages*, 2012. [Online]. Available: https://www.researchgate.net/profile/Mounir_Zrigui/publication/233746674_Proposal_of_a_method_of_enriching_queries_by_statistical_analysis_to_search_for_information_in_Arabic/links/00463514f316c46113000000.pdf#page=54 [Accessed: 17-Des-2019]

[4] Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of sentiment reviews using n-gram machine learning approach." Expert Systems with Applications 57 (2016): 117-126. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S095741741630118X [Accessed: 29-Okt-2020]

[5] Soheily-Khah, Saeid, Pierre-François Marteau, and Nicolas Béchet. "Intrusion detection in network systems through hybrid supervised and unsupervised machine learning process: A case study on the iscx dataset." In 2018 1st International Conference on Data Intelligence and Security (ICDIS), pp. 219-226. [Online]. IEEE, 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8367767 [Accessed: 29-Okt-2020]

[6] Zhou, Shusen, Qingcai Chen, and Xiaolong Wang. "Active deep learning method for semi-supervised sentiment classification." Neurocomputing 120 (2013): 536-546. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0925231213004888 [Accessed: 29-Okt-2020]

[7] Sarkar, Kamal, Mita Nasipuri, and Suranjan Ghose. "Machine learning based keyphrase extraction: Comparing decision trees, naïve Bayes, and artificial neural networks." JIPS 8, no. 4 (2012): 693-712. [Online]. Available: http://jips-k.org/journals/jips/digital-library/manuscript/file/22564/JIPS-2012-8-4-693.pdf [Accessed: 29-Okt-2020]

[8] Dhande, Lina L., and Girish K. Patnaik. "Analyzing sentiment of movie review data using Naive Bayes neural classifier." International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 3, no. 4 (2014): 313-320. [Online]. Available: http://student.blog.dinus.ac.id/windalistyaningsih/wp-content/uploads/sites/316/2017/07/IJETTCS-2014-08-25-138.pdf [Accessed: 29-Okt-2020]

[9] Khairnar, Jayashri, and Mayura Kinikar. "Machine learning algorithms for opinion mining and sentiment classification." International Journal of Scientific and Research Publications 3, no. 6 (2013): 1-6. [Online]. Available: http://www.ijcst.org/Volume4/Issue6/p12_4_6.pdf [Accessed: 29-Okt-2020]

[10] Saif, Hassan, Yulan He, Miriam Fernandez, and Harith Alani. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management 52, no. 1 (2016): 5-19. [Online]. Available: https://publications.aston.ac.uk/id/eprint/25812/1/Contextual_semantics_for_sentiment_analysis_of_Twitter.pdf [Accessed: 29-Okt-2020]

[11] Dey, Lopamudra, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, and Sweta Tiwari. "Sentiment analysis of review datasets using naive bayes and k-nn classifier." arXiv preprint arXiv:1610.09982 (2016). [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1610/1610.09982.pdf [Accessed: 29-Okt-2020]

[12] Mukherjee, Saurabh, and Neelam Sharma. "Intrusion detection using naive Bayes classifier with feature reduction." *Procedia Technology* 4,pp. 119-128, 2012 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212017312002964 [Accessed: 17-Des-2019]

[13] Calders, T., & Verwer, S. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2), pp. 277-292, 2010. [Online]. Available: https://link.springer.com/article/10.1007/s10618-010-0190-x [Accessed: 20-Des -2019]

[14] Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. International Journal of Engineering and Emerging Technology, 2(1), pp.53-57, 2017. [Online]. Available: https://ojs.unud.ac.id/index.php/ijeet/article/view/34457 [Accessed: 17-Des-2019]

[15] Carlin, B. P., & Louis, T. A. Bayes and empirical Bayes methods for data analysis. Chapman and Hall/CRC, 2010. [Online]. Available: https://link.springer.com/article/10.1023/A:1018577817064 [Accessed: 17-Des-2019]

[16] Zaidi, N. A., Cerquides, J., Carman, M. J., & Webb, G. I. Alleviating naive Bayes attribute independence assumption by attribute weighting. The Journal of Machine Learning Research, 14(1), pp.1947-1988, 2013. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/2567709.2567725 [Accessed: 17-Des-2019]

[17] Pandey, U. K., & Pal, S. Data Mining: A prediction of performer or underperformer using classification. arXiv preprint arXiv:1104.4163,2011. [Online]. Available: https://arxiv.org/abs/1104.4163 [Accessed: 17-Des-2019]

[18] Calders, T., & Verwer, S. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2), pp.277-292, 2010. [Online]. Available: https://link.springer.com/article/10.1007/s10618-010-0190-x [Accessed: 20-Des -2019]

[19] Liu, B. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1),pp. 1-167, 2012. [Online]. Available: https://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016 [Accessed: 17-Des-2019]

[20] Zhu, W., Zeng, N., & Wang, N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland, 19,pp. 67, 2010. [Online]. Available: https://www.lexjansen.com/nesug/nesug10/hl/hl07.pdf [Accessed: 20-Des -2019]

[21] Ho, C. Y., Lai, Y. C., Chen, I. W., Wang, F. Y., & Tai, W. H. Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems. IEEE Communications Magazine, 50(3), pp. 146-154, 2012. [Online]. Available: https://ir.nctu.edu.tw/bitstream/11536/15580/1/000301198700019.pdf [Accessed: 20-Des -2019]

[22] Bellazzi, R., Ferrazzi, F., & Sacchi, L. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(5), pp. 416-430, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.23 [Accessed: 17-Des-2019]

[23] Caleffi, P. M. The'hashtag': a new word or a new rule?. SKASE journal of theoretical linguistics, 12(2), 2015. [Online]. Available: http://www.skase.sk/Volumes/JTL28/pdf_doc/05.pdf [Accessed: 17-Des-2019]

[24] Bruns, A., Weller, K., Borra, E., & Rieder, B. Programmed method: Developing a toolset for capturing and analyzing tweets. Aslib Journal of Information Management, 2014. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/AJIM-09-2013-0094/full/html [Accessed: 20-Des-2019]