Bit-Tech

Vol.2, No.1, Oktober 2019
Available online at: http://jurnal.kdi.or.id/index.php/bt

Extraction Opinion of Social Media in Higher Education Using Sentiment Analysis

Thomas E Tarigan [1)*], **Robby C Buwono**[2)], **Sri Redjeki**[3)]

[1)2)3)]*STMIK AKAKOM*
Jln Raya Janti No 143, Yogyakarta, Indonesia
[1)]tarigan@akakom.ac.id
[2)]robby@akakom.ac.id
[3)]dzeky@akakom.ac.id

Article history:

Received 30 September 2019;
Revised 10 October 2019;
Accepted 16 October 2019;
Available online 30 October 2019

Keywords: {use 4-6 keywords}

NBC
Opinion Mining
Sentiment Analysis
Twitter
University

Abstract

The purpose of this research is to extract social media Twitter opinion on a tertiary institution using sentiment analysis. The results of sentiment analysis will provide input to universities as a form of evaluation of management performance in managing institutions. Sentiment analysis generated using the Naïve Bayes Classifier method which is classified into 4 classes: positive, normal, negative and unknown. This study uses 1000 data tweets used for training data needs. The data is classified manually to determine the sentiment of the tweet. Then 20 tweet data is used for testing. The results of this study produce a system that can classify sentiments automatically with 75% test results for sentiment, some obstacles in processing real-time tweets such as duplicate tweets (spam tweets), Indonesian structures that are quite complex and diverse.

I. Introduction

Sentiment analysis is the process of understanding, extracting and processing textual data automatically to get sentiment information contained in an opinion sentence. Sentiment analysis is done to see an opinion or tendency of opinion on a problem or object by someone, whether they tend to have negative views or positive opinions. One example of using sentiment analysis in the real world is the identification of market trends and market opinion on an object of goods. The popularity of social networking media has continued to increase in recent years. Social networking media like Twitter, Facebook, and Youtube are some of the most popular communication media tools that exist among internet users today [1]. Therefore, social networking media is widely used by universities to provide information about tertiary institutions. Twitter users in Indonesia in 2018 increased 11% compared to the previous year. This is the increasing number of social media users in Indonesia using Twitter to write their opinions directly on social media. Twitter has continued to increase in number of users since its appearance in 2006. Statistics show that Twitter has more than 1 billion users, of which 313 million are active users. Many researchers have given more attention to social media. Data analysis using social media is not easy because of its incompleteness and dynamic nature [2] [3].

Twitter is the fastest growing social network since 2006. Twitter is a real-time microblogging service by providing only 140 short characters but can provide insights or meaning enough [4] [5]. The existence of Twitter which has been widely used by various levels of society can be seen as a good reflection where the existence of Twitter can present what is a trend of conversation and what is interesting to be discussed.

Research that discusses the use of social media for the extraction of opinion of a tertiary institution, among others, by Imam Fahrur et al. In this study an opinion mining system was developed to analyze public opinion in tertiary institutions. In the subjectivity and target detection subprocesses, Part-of-Speech (POS) Tagging is used using Hidden Makov Model (HMM). In the POS Tagging process results are then applied rules to find out whether a document is included as an opinion or not, as well as to find out which part of the sentence is an object that is the target of opinion. Documents that are recognized as opinions are then classified into negative and positive opinions (subprocess opinion orientation) using the Naïve Bayes Classifier (NBC). Research on the use of NBC as a method of text classification

ISSN 2622-2728 (online) 2622-271X (print)
doi:

has been carried out by SM Kamaruzzaman and Chowdury Mofizur Rahman [6] and Ashraf M Kibriya et.al. [7] in 2004. From the testing process qualitatively stated that the text can be classified with high accuracy.

This study analyzes and sentiments classification of the social media of private universities in IT in Yogyakarta. The method used in the classification of sentiment categories is Naïve Bayes Classifier. The object to be classified is not at the document level but a word in a Twitter sentence, and classifies whether the tweet is positive, neutral, negative or unknown.

II. Literature Review

Some writings on social media to make estimates or prediction based on text are currently mostly done by social and exact researchers to get information from public areas quickly with a very large amount of data. Big data, one of the emerging fields of science, has provided a way to continue to explore research in social media. One of the most explored social media is Twitter. Some studies are used as a reference for researchers [8] who develop a system for opinion extraction via Twitter to see public opinion on a college. Research [9] titled Sentiment Analysis and Category Classification of Public Figure on Twitter uses the Support Vector Machine method to classify sentiments. Other research [10] also discusses sentiment analysis for figures in Indonesia using SVM based on Cloud Computing. Another article [11] conducted a sentiment analysis used to review restaurants using Indonesian text by utilizing genetic algorithms to improve the performance of the Naïve Bayes Classifier and was found to be able to improve the accuracy of the results of the classification of research results.

A. Sentiment Analysis

Sentiment Analysis or opinion mining is the process of understanding, extracting and processing textual data automatically to get the sentiment information contained in an opinion sentence. Sentiment analysis is done to see an opinion or tendency of opinion on a problem or object by someone, whether they tend to have negative views or positive opinions [12]. In the opinion of Zaqisyah (2012). Sentiment analysis is a process of extracting, processing data and understanding textual automatically so that it can produce sentiment information that depends on an opinion sentence. Thus sentiment analysis can be used to see the tendency of opinion on a problem or object by someone, whether they tend to have a positive or negative outlook. Based on these opinions, it can be said that sentiment analysis is an activity of analyzing one's opinions, opinions, attitudes or emotions about a particular product, topic or problem so that it can be known as positive, negative or neutral sentiments.

B. Naïve Bayes Classifier

Bayes is a simple probabilistic based prediction technique based on applying the Bayes theorem (or Bayes rule) with strong or naïve assumptions of independence. In other words, in Naïve Bayes, the model used is an "independent feature model". In Bayes (especially Naïve Bayes), the purpose of strong independence of features is that a feature in a data is not related to the presence or absence of other features in the same data [14].

In the Naïve Bayes Classifier algorithm each document is represented by the attribute pair "x1, x2, x3, ... xn" where x1 is the first word, x2 is the second word and so on. Whereas V is the set of headline categories. At the time of classification the algorithm will look for the highest probability of all categories of documents tested (V_MAP), where the equation is as follows:

$$V_{MAP} = \underset{V\,j\;e\,V}{\arg\ max} \frac{P(x_1, x_{2,}x_{3,\dots}x_n|Vj)P(Vj)}{P(x_{1,}x_{2,}x_{3,\dots}x_n|Vj)} \dots\dots\dots\dots\dots\dots(1)$$

For P (x1, x2, x3, ... xn) the values are constant for all categories (Vj) so the equation can be written as follows:

$$V_{MAP} = \underset{V\,j\;e\,V}{\arg max}\ P\big(x_1, x_{2,}x_{3,\dots}x_n\big|Vj\big)P(Vj) \dots\dots\dots\dots\dots\dots(2)$$

The equation above can be simplified as follows:

$$V_{MAP} = \underset{V\,j\;e\,V}{\arg max} \prod_{i=1}^{n} P(x_1|Vj)P(Vj) \dots\dots\dots\dots\dots\dots(3)$$

Variabel Information :
Vj = Headline Category
J = 1, 2, 3,…n.
j1 = Headline category for positif sentiment

j2 = Headline category for negative sentiment
j3 = Headline category for netral sentiment
P(xi|Vj) = Probability of xi in category Vj
P(Vj) = Probability from Vj

For P(Vj) and P(xi|Vj) calculated during training where the equation is as follows:

$$P(Vj) = \frac{|docs\ j|}{|contoh|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

If P(Vj) It has been determined then count the number of documents for each category j and the number of documents from all categories using formula 4. above.

$$P(xi|Vj) = \frac{nk+1}{nk+|kosakata|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

If P(xi|Vj) has been determined, then count the number of times the appearance of each word plus 1 and the number of times the appearance of words from each category using the formula 5. above.

Variabel informatio:
|docs j| = Documents count of every category j
|contoh| = Documents count from all category
nk = the number of times each word appears
n = the number of times the word appears in each category
|kosakata| = the sum of all words from all categories

In the classification using Naïve Bayes is divided into 2 processes, namely the training and testing process. The training process is used to produce a sentiment analysis model which will later be used as a reference for classifying sentiments with new testing data or raw data.

III. METHODS

This research uses Twitter data with name keywords and things related to tertiary institutions that will extract public opinion. The tertiary institution which is the object of this research is the IT field in Yogyakarta. The amount of Twitter data collected is 1000 Twitter data that has been clean and given manual labeling class. Giving labeling class is done by using basic word data dictionary which has the meaning of positive, neutral and negative sentiments. The flow of the research mechanism is shown in Figure 1 in the form of a system block diagram.

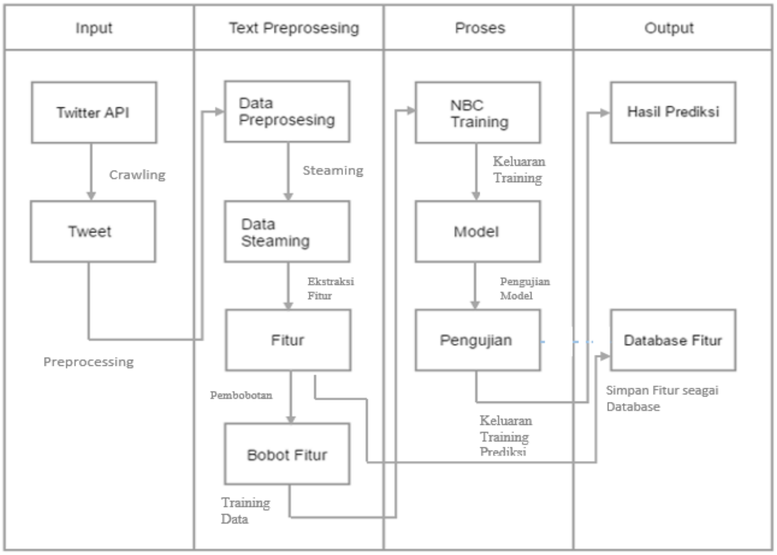


Figure 1. Diagram Block System

In Figure 1, there are 4 blocks that show stages, namely input consisting of 2 sub-stages, namely using the Twitter API to crawl data from existing tweets. Keyword uses words related to tertiary institutions which will be the object of research. In the next stage is preprocessing in which there are 4 sub stages, namely preprocessing data, steaming process, feature extraction and weighting of each feature and labeling class. Data generated at this stage will be used for the process stage. In the third stage or this process begins training and testing using the classification method, namely Naïve Bayes Classifier. There are 3 classifications in this system: positive, neutral and negative sentiments or opinions. The last stage is the stage of producing output in the form of predictions, where users can enter any tweet or opinion and the results of their classification predictions will be obtained. A description of the relationship between the system and the actors developed in this study is shown in Figure 2. There are 2 actors namely engineer and Twitter API as external factors while the internal factors of the system there are 4 main functionalities namely Twitter data load, preprocessing, process using Naïve Bayes Classifier algorithm and results research in the form of classification results sentiment analysis into 3 classes, namely positive, neutral and negative.

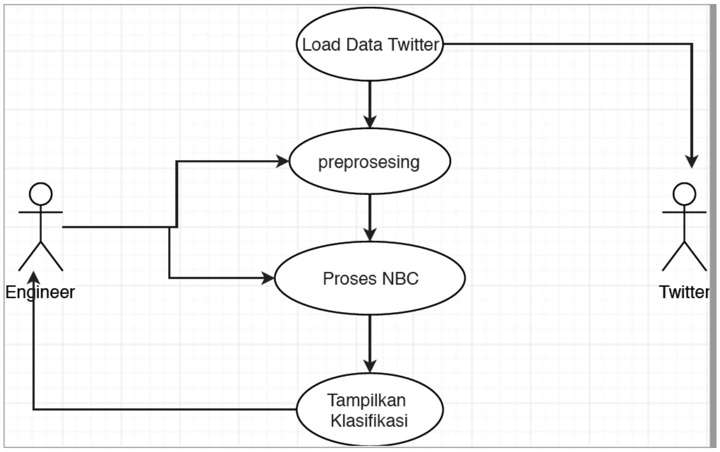


Figure 2. Usecase Diagram of Classification System

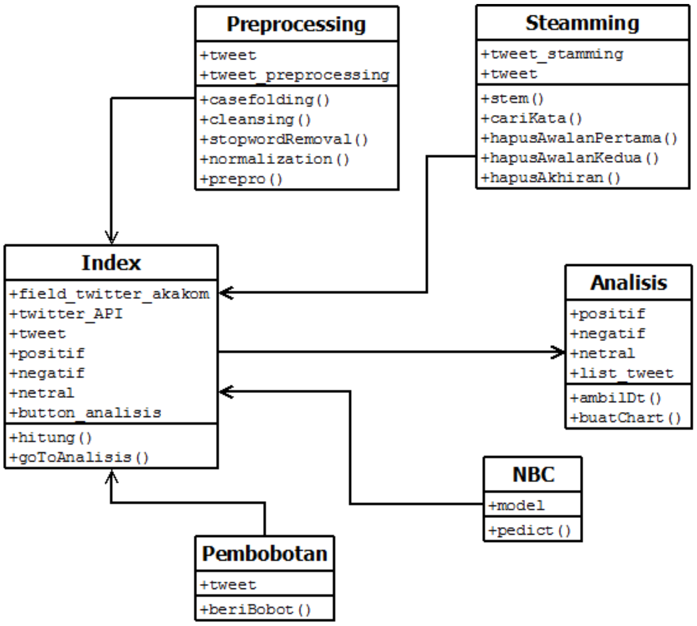


Figure 3. Class Diagram System

In Figure 3 shows the system class diagram developed in this study. There are 6 classes used on the system, namely preprocessing class, steaming class, class indexing, analysis class, weight class and NBC class. The system on the client side is used to present data into several parts, namely the percentage of sentiment and tweet data, after getting the sentiment percentage data and tweet data, the client application will present it in the form of a pie chart. On the server side there are several types of classes, namely the index class which is the main class of software, and there are preprocessing, steaming, and weighting classes.

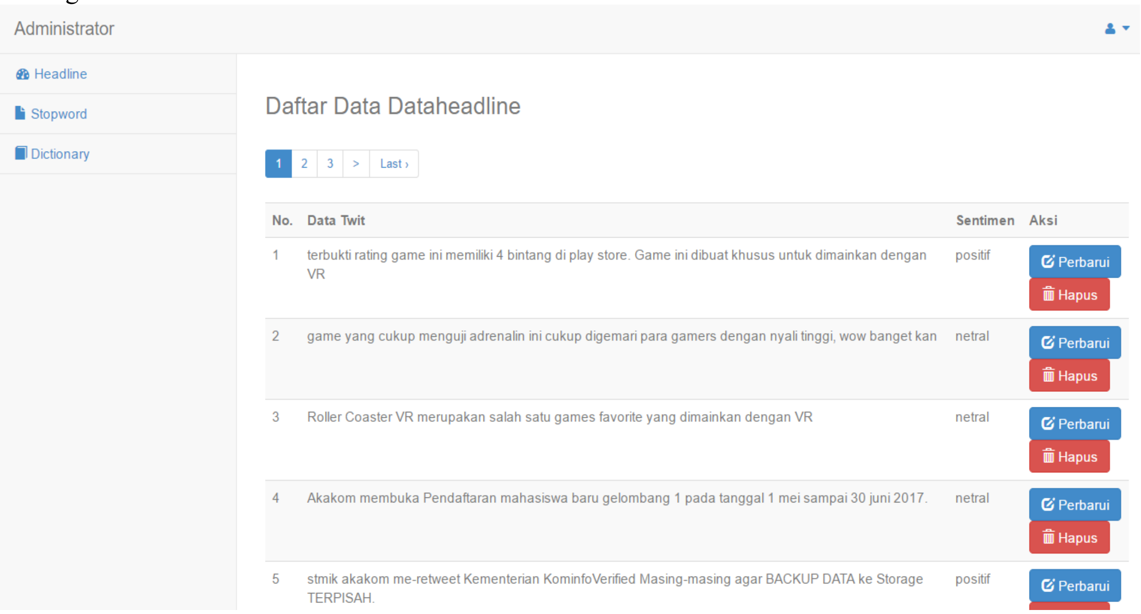
IV. RESULTS

This section will discuss the results of training and testing of the model used, the Naïve Bayes Classifier (NBC). The amount of training data used is 1000 data tweets that have been done preprocessing.

Table 1. Tweet and Classification

No.	Kata	Kategori
1.	Fasilitas kampus lengkap	Positif
2.	Dosen ramah semua	Positif
3.	Mahasiswa jogja	Netral
4.	Kuliah padat	Netral
5.	Dosen sering terlambat	Negatif
6.	Kuliah susah sekali	Negatif

Table 1 is an example of a clean tweet and labeled manually. Each training data will be calculated its probability value against the possibility of all three classes using Bayesian. Twitter data that has been obtained is displayed in the dashboard figure 4.

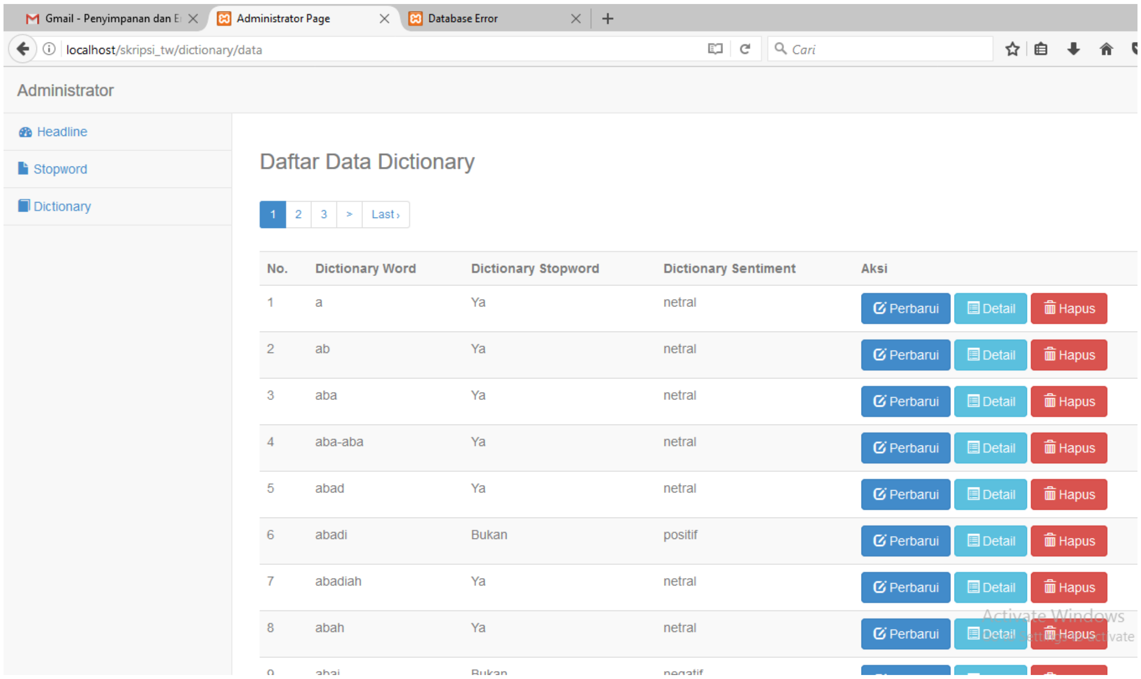


Daftar Data Dataheadline

No.	Data Twit	Sentimen	Aksi
1	terbukti rating game ini memiliki 4 bintang di play store. Game ini dibuat khusus untuk dimainkan dengan VR	positif	Perbarui Hapus
2	game yang cukup menguji adrenalin ini cukup digemari para gamers dengan nyali tinggi, wow banget kan	netral	Perbarui Hapus
3	Roller Coaster VR merupakan salah satu games favorite yang dimainkan dengan VR	netral	Perbarui Hapus
4	Akakom membuka Pendaftaran mahasiswa baru gelombang 1 pada tanggal 1 mei sampai 30 juni 2017.	netral	Perbarui Hapus
5	stmik akakom me-retweet Kementerian KominfoVerified Masing-masing agar BACKUP DATA ke Storage TERPISAH.	positif	Perbarui

Figure 4. Dashboard Data

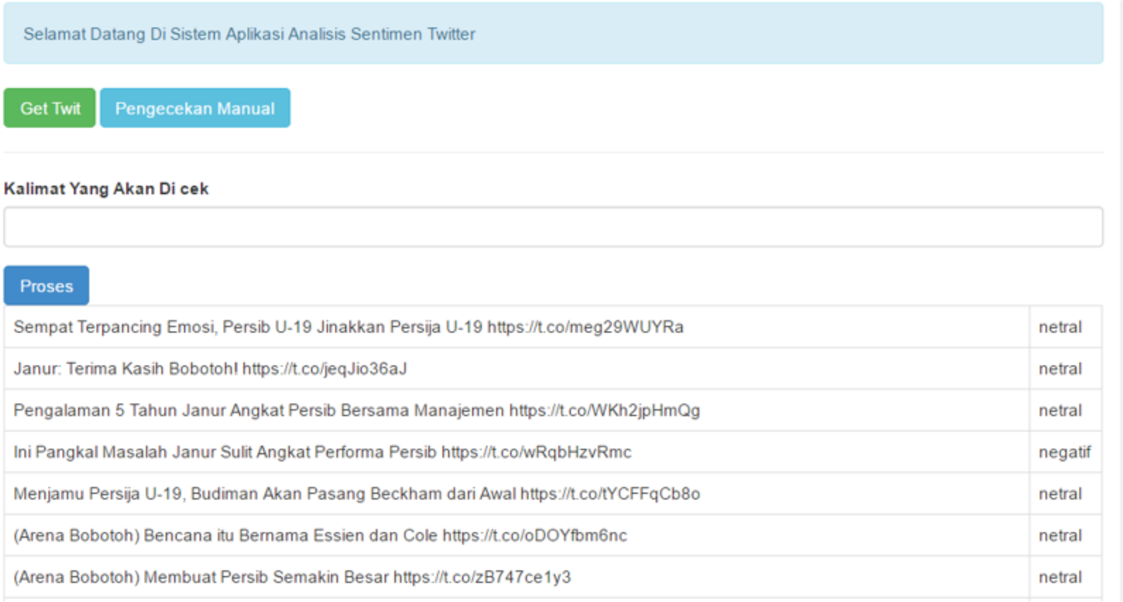
The data dictionary is also displayed in Figure 5. The user can edit the data if there is a change in the data dictionary. This menu is to facilitate the management of data dictionaries that have a tendency to increase.



Daftar Data Dictionary

No.	Dictionary Word	Dictionary Stopword	Dictionary Sentiment	Aksi
1	a	Ya	netral	Perbarui Detail Hapus
2	ab	Ya	netral	Perbarui Detail Hapus
3	aba	Ya	netral	Perbarui Detail Hapus
4	aba-aba	Ya	netral	Perbarui Detail Hapus
5	abad	Ya	netral	Perbarui Detail Hapus
6	abadi	Bukan	positif	Perbarui Detail Hapus
7	abadiah	Ya	netral	Perbarui Detail Hapus
8	abah	Ya	netral	Perbarui Detail Hapus
9	abai	Bukan	negatif	

Figure 5. Dashboard Data Dictionary



Selamat Datang Di Sistem Aplikasi Analisis Sentimen Twitter

Kalimat Yang Akan Di cek

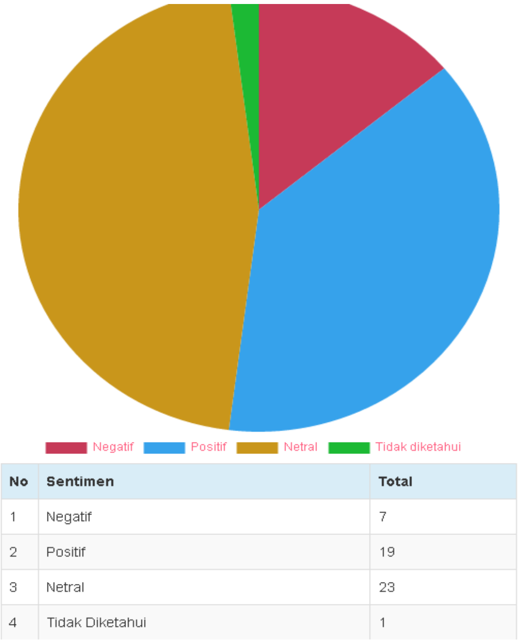
Sempat Terpancing Emosi, Persib U-19 Jinakkan Persija U-19 https://t.co/meg29WUYRa	netral
Janur: Terima Kasih Bobotoh! https://t.co/jeqJio36aJ	netral
Pengalaman 5 Tahun Janur Angkat Persib Bersama Manajemen https://t.co/WKh2jpHmQg	netral
Ini Pangkal Masalah Janur Sulit Angkat Performa Persib https://t.co/wRqbHzvRmc	negatif
Menjamu Persija U-19, Budiman Akan Pasang Beckham dari Awal https://t.co/tYCFFqCb8o	netral
(Arena Bobotoh) Bencana itu Bernama Essien dan Cole https://t.co/oDOYfbm6nc	netral
(Arena Bobotoh) Membuat Persib Semakin Besar https://t.co/zB747ce1y3	netral

Figure 6. Dashboard Testing Tweet

Figure 6 is used to see the results of the tweet classification input through the tweet testing dashboard. The classification results will be displayed in the form of a list of tweets and the type of classification.

V. Discussion

The final result of the classification of testing data is displayed in the form of a circular graph which is divided into 4 areas: positive, neutral, negative and unknown. This graph is shown in Figure 7.



No	Sentimen	Total
1	Negatif	7
2	Positif	19
3	Netral	23
4	Tidak Diketahui	1

Gambar 7. Grafik Klasifikasi

By using 1000 training data and testing data as many as 20 then manually tested and compared with the results of the classification conducted by the system. Obtained as many as 5 testing data and 15 correct testing data, then from these results obtained an accuracy value of 75%.

Table 2. Result of Data Testing

No.	**Data Tweet Testing**	**Sistem**	**Dictionary**	**Hasil**	**Cek Dictionary**	**Dataset**
1	Kebanggaan kita yang terbesar bukan karena tidak pernah gagal, tetapi bangkit kembali setiap kita jatuh - akakom	Positif	Netral	1	bangga besar gagal tetapi bangkit setiap jatuh	0 5 2
2	Kepuasan terletak pada usaha, bukan pada hasil. Berusaha dengan keras adalah kemenangan yang hakiki - akakom	Positif	Positif	0	puas usaha bukan hasil keras menang hakiki	4 3 0
3	Kita berdoa jika kesusahan dan membutuhkan sesuatu, mestinya kita juga berdoa dalam kegembiraan besar dan rezeki melimpah - akakom	Positif	Netral	1	doa susah butuh gembira besar rezeki melimpah	2 4 1
4	Tiadanya keyakinanlah yang membuat orang takut menghadapi tantangan, dan saya percaya pada diri saya sendiri - akakom	Positif	Netral	1	yakin takut hadapi tantangan percaya diri sendiri	1 5 1
5	Pendidikan merupakan senjata yang memiliki kekuatan untuk mengubah dunia - akakom -	Netral	Netral	0	didik senjata kuat ubah dunia	1 4 0
6	Pendidikan yang sejati akan melahirkan harapan baru - akakom -	Positif	Positif	0	didik sejati lahir harapan baru	3 2 0
7	Saya bukan orang pintar, namun saya telah mengenal, menerima, dan memikirkan banyak sekali pertanyaan - akakom -	Positif	Positif	0	bukan pintar kenal terima mikir pertanyaan	3 2 1

8	Pendidikan bukanlah persiapan untuk hidup, ia adalah hidup itu sendiri - akakom -	Positif	Positif	0	didik bukan siap sendiri	2 1 1
9	Ceritakan kepadaku, maka aku akan lupa. Ajarkan aku, mungkin aku bisa mengingatnya.Ajak dan libatkan lah aku, maka aku akan belajar-akakom	Positif	Netral	1	cerita lupa ajar ingat libat belajar	0 5 1
10	Kulit dari pendidikan itu memang pahit, namun buahnya sangatlah manis dan aromanya wangi - akakom -	Positif	Positif	0	Kulit didik pahit buah manis wangi	3 2 1
11	Pendidikan adalah kunci untuk membuka pintu emas kebebasan - akakom -	Netral	Netral	0	didik kunci buka pintu emas bebas	1 5 0
12	Tujuan utama dari pendidikan adalah mengubah JENDELA menjadi PINTU - akakom -	Netral	Netral	0	Tuju utama didik ubah jendela pintu	1 5 0
13	Filosofi dari pendidikan saat ini akan menjadi filosofi pemerintahan dimasa yang akan datang - akakom -	Netral	Netral	0	Filosofi didik saat jadi masa datang	1 5 0
14	Belajarlah dari masa lalu jika kita ingin mendefinisikan masa depan - akakom -	Netral	Netral	0	Belajar masa definisi depan	0 4 0
15	Dengan mencari dan Berspekulasi maka kita akan belajar dan mendapatkan hal-hal yang baru - akakom -	Netral	Netral	0	spekulasi belajar dapat hal baru	0 5 0
16	Pendidikan seseorang takkan sempurna sampai kematian mendatanginya - akakom -	Negatif	Netral	1	didik sempurna sampai mati datangi	1 3 1
17	Saya tidak berbicara dengan kata mungkin - akakom -	Negatif	Netral	1	tidak bicara dengan kata mungkin	0 4 1
18	Persiapkan hari ini untuk keinginan hari esok - akakom -	Netral	Netral	0	siap untuk ingin esok	0 4 0
19	Kesenangan dalam sebuah pekerjaan membuat kesempurnaan pada hasil yang dicapai - akakom -	Positif	Positif	0	senang kerja sempurna hasil capai	3 2 0
20	Yang membuatku terus berkembang adalah tujuan-tujuan hidupku - akakom -	Netral	Netral	0	buat kembang tujuan hidup	0 4 0

VI. Conclusions

Some things that are made conclusions from the results of research, including:

1. The Naïve Bayes Classifier method can be used as an algorithm that is good enough to do sentiment analysis
2. The accuracy of the classification testing using the Naive Bayes Classifier method is 75%.
3. Indonesian sentences with more complex sentence structure and variations require a lot of training data to get good results.
4. Public opinion about a university can be used as a reference for universities to improve management and performance.

References

[1] Dave Chaffey, Fiona Ellis-Chadwick, Digital Marketing (6th Edition), Pearson, 2016

[2] Wang, P., Xu, B., Wu, Y., & Zhou, X., (2015), Link prediction in social networks: the state-of-the-art., Science China-Information Science, Vol. 58 011101:1–011101:38., link.springer.com.

[3] Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS, (2015), Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol 11(10): e1004513. https://doi.org/10.1371/journal.pcbi.1004513.

[4] Nan Li, D.D, & Wu, (2010), Decision Support System 48, Page 354-368, Elsevier.

[5] Li, C., Bai, J., Zhang, L., Tang, H., & Luo, Y. (2019). Opinion community detection and opinion leader detection based on text information and network topology in cloud environment. Information Sciences. doi:10.1016/j.ins.2019.06.060

[6] Kamaruzaman, S.M., Chowdhury M.R. 2004. *Text Categorization using Association Rule and Naive Bayes Classifier*. Asian Journal of Information Technology, Vol. 3, No.9, pp 657-665, Sep. 2004

[7] Kibriya Ashraf M., Frank Eibe, Pfahringer Bernhard, Holmes Geoffrey . 2004. *Multinomial Naïve Bayes for Text Categorization Revisited*. Australian joint conference on artificial intelligence No 17.

[8] Rozi, Imam Fahrur, Sholeh Hadi Pramono, and Erfan Achmad Dahlan. "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi." *Jurnal EECCIS* 6.1 (2013): 37-43

[9] Hidayatullah, Ahmad Fathan, and Azhari SN Azhari. "Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada twitter." *Seminar Nasional Informatika (SEMNASIF)*. Vol. 1. No. 1. 2015.

[10] Maulana, Rizky, and Sri Redjeki. "Analisis Sentimen Pengguna Twitter Menggunakan Metode Support Vector Machine Berbasis Cloud Computing." *Jurnal TAM (Technology Acceptance Model)* 6 (2017): 23-28.

[11] Muthia, Dinda Ayu. "Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes." *Jurnal Ilmu Pengetahuan dan Teknologi Komputer* 2.2 (2017): 39-45.

[12] Liu, B. 2012. Sentiment Analysis And Opinion Mining. Chicago, United Stade of America. Morgan and Claypool Publishers. Website : https://books.google.co.id/books?id=Gt8g72e6MuEC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false diakses 2 November 2016

[13] Zaqisyah. 2012. Optimasi Akurasi Analisis Sentimen Pada Posting Twitter Menggunakan Metode Naïve Bayes dan Steamming. Tugas Akhir Skripsi. Teknik Informatika Ilmu Komputer Universitas Komputer Indonesia. Bandung. Website : http://elib.unikom.ac.id/gdl.php?mod=browse&op=read&id=jbptunikompp-gdl-zaqisyahni-35251&q=zaqisyah%202012 diakses 12 November 2016.

[14] Eko Prasetyo. 2012. “Data Mining : Konsep dan Aplikasi Menggunakan MATLAB”. Yogyakarta : ANDI Yogyakarta.