The Fifth Information Systems International Conference 2019

Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm

Veny Amilia Fitri*, Rachmadita Andreswari, Muhammad Azani Hasibuan

Information Systems Department School of Industrial Engineering, Telkom University, Bandung, 40257, Indonesia

Abstract

People use social media as a means to express their thoughts, interests, and opinions on various things. Thousands of submissions occur every day on every social media. Everyone can express their opinions through social media freely. These opinions contain positive, negative and neutral sentiments on a topic. The case study taken by researchers is the Anti-LGBT campaign in Indonesia. The case was taken because the Anti-LGBT campaign was widely discussed by the Indonesian people on Twitter's social media. If you want to know the tendency of public comments on the Anti-LGBT campaign in Indonesia, is it positive, negative, or neutral, then a sentiment analysis is conducted. The algorithm used in conducting sentiment analysis is Naïve Bayes because it has a high degree of accuracy in classifying sentiment analysis. The stages in conducting sentiment analysis in this study are preprocessing data, processing data, classification, and evaluation. The sentiment analysis obtained in this study shows that Twitter users in Indonesia give more neutral comments. In this study, an accuracy of 86.43% was obtained from testing data using Naïve Bayes Algorithm in RapidMiner tools, where the accuracy is higher than the other algorithms, Decision Tree and Random Forest which is 82.91%.

Keywords: Twitter; Naïve Bayes; Anti-LGBT campaign; sentiment analysis

* Corresponding author. Tel.: +62-82-257-237-819.
  E-mail address: venyamilia.fitri@gmail.com

1. Introduction

The rapid development of social media, more and more people are writing their opinions on something. Therefore, this encourages research to use Twitter social media as a data source which will be analyzed in the form of sentiment analysis which is a process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in an opinion sentence [1].

Communities in the world in general and Indonesia, in particular, use social media as a means to express their thoughts, interests, and opinions on various things. The case study taken by researchers is the Anti-LGBT campaign. Taking case studies of the Anti-LGBT campaign for sentiment analysis because this movement has been talked about a lot by Indonesian people on social media. Based on these facts, it can be concluded that the Anti-LGBT campaign attracted a lot of people's interest to be discussed. This is evidenced by the many comments on social media, especially Twitter by the Indonesian people against the Anti-LGBT campaign. Based on the comments on Twitter, sentiment analysis can be carried out.

One way to do sentiment analysis can be done using data from social media. Thousands of submissions occur every day on every social media. Everyone can express their opinions through social media freely. These opinions contain positive, negative and neutral sentiments on a topic. Positive sentiments express a good opinion on a context, negative sentiments express bad opinions in a context, while neutral sentiments state things that do not favor good or bad. The algorithm used in conducting sentiment analysis in this study is Naïve Bayes, Decision Tree, and Random Forest.

2. Literature review

2.1. LGBT

LGBT stands for Lesbian, Gay, Bisexual, Transgender. LGBT is a term used since the 1990s, replacing the phrase gay community because the term is considered more representative of the groups mentioned [2]. Lesbians are homosexuality among women. Gay is homosexuality among men. Bisexuality is a state of feeling equally attracted to both sexes, women, and men [3]. Lesbian, gay, bisexual and transgender (LGBT) behavior is still a sensitive problem in Indonesia. This behavior continues to be polemic because there are a number of parties who want LGBT people to be able to live normally on the basis of human rights and not be subjected to persecution. But the LGBT campaign in Indonesia - the country with the largest Muslim population in the world - is certainly not easy. Many parties, especially Muslims, reject this LGBT behavior because it is deemed deviant and forbidden by religion [4].

2.2. Social network

The social network is a media for socializing with one another and done online that allows humans to interact with each other without being limited by space and time. An individual can be affected in their social network environment by events that occur around their surroundings. The first social network began with the launch of SixDegrees.com in 1997 where users or users can create profiles and register their friends while exploring friendships first began in 1998. In the 2000s, many social networks began to emerge starting with Friendster, Match. com, MySpace, Twitter, to Facebook which has a feature for developers outside Facebook to build applications that allow users to personalize their profile [5].

Twitter is a microblogging service that was officially released on July 13, 2006 [6]. Twitter's main activity is to post something short (tweet) via the web or mobile. The maximum length of a tweet is 280 characters. Twitter is an almost unlimited source used in text classification. There are many characteristics of Twitter tweets [7]. The message on Twitter has many unique attributes, which distinguishes it from other social media:

- Twitter has a maximum character length of 280 characters.
- Twitter provides data that can be accessed freely using the Twitter API, making it easier to collect a large number of tweets.
- Language models - Twitter users post messages through many different media. The frequency of spelling, slang, and abbreviations is higher than other social media.
- Twitter users send short messages about various topics tailored to a particular topic and that applies globally.

2.3. Text mining

As one of the main fields in data mining, Text Mining aims to find previously unknown but potentially useful knowledge from unstructured or semi-structured text data [8]. Text mining also faces problems such as large amounts of data, high dimensions, data and structures that are constantly changing, and data "noise." Unlike data mining, which mainly processes structured data, data used by text mining is generally unstructured, or at least semi structured text.

2.3.1. Natural language processing

Natural language processing (NLP) in theory is a computational technique for analyzing and describing text naturally in one or more levels of linguistic analysis with the aim of achieving the ability to understand the essence of a sentence such as humans for example language processing for various tasks and applications. Natural language processing provides both theory and implementation for various applications [9].

2.3.2. Sentiment analysis

Sentiment analysis is a research branch of text mining. The focus of sentiment analysis is to analyze opinions from a text document. In the business world, sentiment analysis is widely used to automatically analyze customer opinions about products and services. Information is data that has been processed in such a way that it is useful for users who need to help make decisions. Information itself can be divided into two, namely facts and opinions. Facts are objective statements about something that has happened and are usually accompanied by evidence, whereas opinions are more subjective in how someone expresses themselves towards everything that happens according to their respective perceptions and assumptions [10]. Sentiment analysis is defined as the task of finding the author's opinion about a particular entity. Sentiment analysis can be based on and assessed at the level of documents, sentences, or words [11].

2.4. State of the art text mining algorithm

Table 1 shows the list of other studies that have been done before. Generally described the contents of the objectives and results of the research paper that are relevant to the case taken by the current researcher.

Table 1. State of The Art Text Mining Algorithm.

Number	Paper Title	Author	Result
1	Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds	Anmol Nayak and Dr. S Natarajan	Naïve Bayes is the most accurate agorithm with a score of 89% followed by 88% SVM and 85% Random Forest.
2	Sentiment Analysis Based on Public Comments on Online Stores in Facebook Social Media (Case Study: Zalora And Berrybenka)	Syahmia Gusriani	Comparisons k-NN, NB and DT obtained classifications that have the best accuracy are Naïve Bayes with an average accuracy of 90.1%.
3	Comparative Analysis of Decision Tree (C4.5) and Naïve Bayes Classifier Algorithms to Classify New Student Admission at University Level	Suyadi, Arief Setyanto, dan Hanif Al Fattah	DT algorithm (C4.5) has an accuracy of 81.84% and Naive Bayes 85.12%

Based on the state-of-the-art table above, it can be seen that to conduct sentiment analysis in determining positive, negative, and neutral trends using the Naïve Bayes algorithm. This can be indicated by the results of accuracy on the Naïve Bayes algorithm better than the other classification algorithms based on the results of previous studies.

3. Research methodology

In this paper, the research methodology is carried out through several stages, including the identification of problems and solutions, the proposed method, and the last is evaluation/conclusion. The steps taken are used to process twitter data to get the results of sentiment analysis using the Naïve Bayes Algorithm. Fig. 1 shows the research methodology in the study.
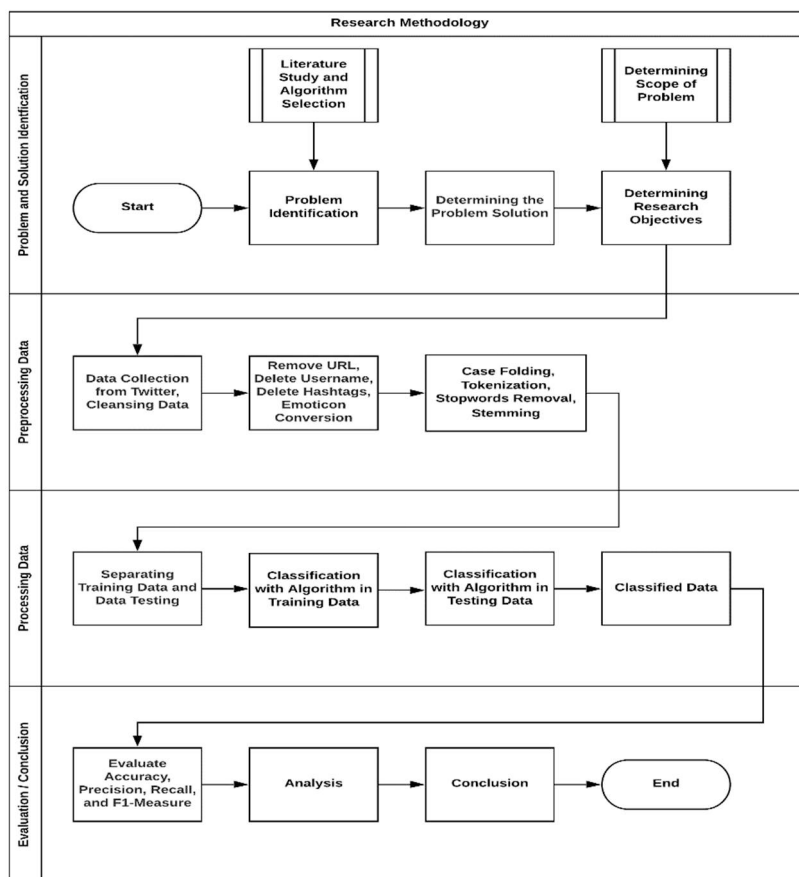

Fig. 1. Research Methodology.

3.1. Problem and solution identification

At the stage of identifying problems and solutions, researchers identify problems with Twitter users who provide comments or comments. At this stage, literature studies are carried out related to the selection of algorithms and how to apply them. Next, determine the solution to the problem. Finally, the purpose of the study is to determine the limits of the problem be studied.

3.2. Preprocessing data

In the proposed method stage, researchers collected data from Twitter. After classification, then preprocessing data, including data cleansing, folding cases, tokenization, stop words removal, and stemming. Next, continue to conduct sentiment analysis using the Naïve Bayes algorithm.

- Data cleansing is cleaning up irrelevant tweet data so that it becomes relevant data.
- Case folding is the process of converting words into the same shape, for example, lowercase or uppercase letters.
- Tokenization is dividing sentences into sections called tokens. Tokens can be formed in words, phrases or other meaningful elements.
- Stopwords removal is the removal of common and often used words but does not have a significant effect in sentences. Twitter messages according to the stopwords list containing stop words in Indonesian such as *dan* (and), *atau* (or), etc.
- Stemming, which is the process of getting the word base by eliminating affixes and suffixes [12]

Next, continue to do sentiment analysis using the Naïve Bayes algorithm.

3.3. Processing data

In the processing data stage, researchers separate training data with testing data. Training data is 75% and data testing is 25%. Then do the classification with the Naïve Bayes algorithm on training data. After that, the classification results from the training data are applied to the testing data. After the data is classified, then perform a performance evaluation.

3.4. Evaluation or conclusion

At the evaluation/conclusion stage, it is done after the data processed by the Naïve Bayes Algorithm produces how many positive, negative and neutral sentiments. To determine the accuracy of the algorithm, evaluation of accuracy, precision, and recall was carried out to produce performance conclusions from the Naïve Bayes algorithm to analyze sentiments on user comments on Twitter. Next, do the analysis and finally conclude the results of sentiment analysis that has been done with the Naïve Bayes algorithm.

In a classifier, there is always something called False Statement. For example, a sentence is declared negative even though the sentence is a positive sentence or a sentence is declared neutral even though the sentence is a positive/negative sentence, as well as a positive sentence that is declared negative (See Table 2). To avoid this we must perform a performance evaluation to find out the False Statement [13].

Table 2. Classification of Performance Evaluation.

Classification Categories	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

How much success of the algorithm is used, it is necessary to check accuracy as seen in Eq. 1:

$$\text{Accuracy} = \frac{TP+T}{TP+FN+FP+T} \tag{1}$$

Precision is the ratio of the relevant items selected to all selected items. Precision can be interpreted as a match between the request for information and the answer to the request (See Eq, 2).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

A recall is defined as the ratio of relevant items selected to the total number of relevant items available. A recall is calculated by Eq. 3 [14]:

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3)$$

But in this study using 3 sentiment, there are positive, negative, and neutral.

4. Result and discussion

The data used in this study is data from social media Twitter with comments on discussions about the Anti-LGBT campaign. the keyword used is using a hashtag that is within the scope of the case discussed. The data used are relevant data that has been cleansed, which is 3744 comments. The data is manually labeled as positive, negative, and neutral sentiments. There are some examples of such tweets manual labeling by different sentiment (See Table 3).

Table 3. Example tweets which are correctly classified.

Tweets	Sentiment
Bebaskan negeri ini dari LGBT	Positive
(Free this country from LGBT)	
Mendingan kampanyekan tolak yg haram seperti LGBT	Positive
(It is better to reject illegitimate campaigns like LGBT)	
Yes Dukung!!	Positive
(Yes Support!!)	
Tolak Anti-LGBT	Negative
(Reject Anti-LGBT)	
Itu hak mereka untuk melakukan penyimpangan	Negative
(It's their right to make a deviation)	
makanya aku tidak demo tolak lgbt	Negative
(therefore I didn't do a demo of rejecting LGBT)	
Prabowo gunakan simbol telunjuk jempol	Neutral
(Prabowo uses the thumbprint symbol)	
Yg kita bahas ini tentang LGBT	Neutral
(What we discussed is about LGBT)	
Kemaren saya diam aja dari pro kontra LGBT	Neutral
(Yesterday I was silent about the pros and cons of LGBT)	

The data used in this study are comments/tweets of Indonesian people on Twitter that are taken using Python, then saved into CSV format files. Keywords for searching comments using hashtags and query searches are related to Anti-LGBT case studies. The data used are data that are relevant to the case studies discussed. The results of relevant data collected were 3744 comments/tweets. The training and testing data ratio is 75:25 where 75% of the data will be used as training data and 25% of the data will be used as testing data. Table 4 presents the distribution of training data and testing data.

Table 4. Data Training and Data Testing Distribution.

	Training Data	Testing Data	Total
Amount of Data	2808	936	3744

Next to manual labeling is compared with the results of sentiment analysis using the Naïve Bayes algorithm on RapidMiner tools. The results are obtained as depicted in Table 5.

Table 5. Result Classification of Performance Evaluation Naïve Bayes Algorithm.

Classification Categories	True Positive	True Negative	True Neutral	Precision	Recall	F1-Measure
Prediction Positive	102	4	53	64.15%	68.46%	66.23%
Prediction Negative	25	4	20	8.16%	36.36%	13.33%
Prediction Neutral	22	3	703	96.57%	90.59%	93.48%

From the results of the data in the table above, data with a total of 936 comments/tweets, obtained 102 comments declared true positive, 4 comments declared true negative, and 703 comments declared true neutral. From that result, So the average percentage each of Recall, Precision and F1-Measure are between 56% and 65%. The Naïve Bayes algorithm testing in this study using RapidMiner tools produced an accuracy of 83.43%. The computational time of this algorithm is about 15 seconds.

Table 6. Result Classification of Performance Evaluation Decision Tree Algorithm.

Classification Categories	True Positive	True Negative	True Neutral	Precision	Recall	F1-Measure
Prediction Positive	0	0	0	0.00%	0.00%	0.00%
Prediction Negative	0	0	0	0.00%	0.00%	0.00%
Prediction Neutral	149	11	775	82.91%	100%	90.66%

From the results of the data in the table above, data with a total of 936 comments/tweets, obtained 0 comments declared true positive, 0 comments declared true negative, and 775 comments declared true neutral. From that result, So the average percentage each of Recall, Precision and F1-Measure are between 27.66% and 33.33%. The Decision Tree algorithm testing in this study using RapidMiner tools produced an accuracy of 82.91% (See Table 6). The computational time of this algorithm is about 13 seconds.

Table 7. Result Classification of Performance Evaluation Random Forest Algorithm.

Classification Categories	True Positive	True Negative	True Neutral	Precision	Recall	F1-Measure
Prediction Positive	0	0	0	0.00%	0.00%	0.00%
Prediction Negative	0	0	0	0.00%	0.00%	0.00%
Prediction Neutral	149	11	775	82.91%	100%	90.66%

From the results of the data in the table above, data with a total of 936 comments/tweets, obtained 0 comments declared true positive, 0 comments declared true negative, and 775 comments declared true neutral. From that result, So the average percentage each of Recall, Precision and F1-Measure are between 27.66% and 33.33%. The Random Forest algorithm testing in this study using RapidMiner tools produced an accuracy of 82.91% (See Table 7). The computational time of this algorithm is about 1.27 minutes.

The result of true negative and true positive is 0 because there is a gap in the amount of neutral data that is far away. So DT and RF fail to predict the class positive and negative. In experiments conducted on data testing, there are still many misspelling and many abbreviations and slang words that cannot be processed when preprocessing data. In addition, this study still cannot process slang words and abbreviations into standard words to facilitate the classification process in Indonesian. As for stopwords removal and stemming on RapidMiner tools researchers conducted custom dictionaries in Indonesian so that they were still not perfect in preprocessing data because the dictionary was still not provided in Indonesian

5. Conclusion

   Based on comment data obtained from twitter about the Anti-LGBT campaign the tendency for comments to be submitted contained neutral comments. Twitter users are neutral about the issue. From 936 testing data, there are 703 comments with the neutral sentiment, then 102 comments with positive sentiment, and 4 comments with negative sentiment. Here it can be concluded that social media users in Indonesia are neutral towards the anti-LGBT campaign, but more are supporting the Anti-LGBT campaign than those who refused. Based on the result of the accuracy of the Naïve Bayes algorithm in the amount of 83.43%, that higher than the accuracy of Decision Tree Algorithm and Random Forest Algorithm it can be concluded that the Naïve Bayes algorithm is very well used to conduct sentiment analysis in the case of Anti-LGBT campaign on Twitter social media.

Acknowledgements

References

[1]	Fahrur Rozi, Imam, Sholeh Hadi Pramono, and Erfan Achmad Dahlan. (2012) "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi [Title in English: *The Implementation of Opinion Mining to Extract Public Opinion Data in Higher Education*]." *Jurnal EECCIS* **37**.
[2]	Yansyah, Roby, and Rahayu. (2018) "Globalisasi Lesbian, Gay, Biseksual, Dan Transgender (LGBT): Perspektif HAM dan Agama Dalam Lingkup Hukum Di Indonesia [Title in English: *The Globalization of Lesbian, Gay, Bisexual, and Transgender (LGBT): The perspective of Human Rights, Religion in The Aspect of Laws in Indonesia*]." *Jurnal Law Reform*. pp. 132-146
[3]	Chaplin, J.P. (2006) "Kamus Lengkap Psikologi [Title in English: *Complete Dictionary of Psychology*]", Jakarta, Indonesia, PT. RajaGrafindo Persada.
[4]	Wardah, Fathiyah. (2018) *VOA Indonesia*. Available from: https://www.voaindonesia.com.
[5]	M. Boyd, Danah, and Nicole B. Ellison. (2010) "Social Network Sites: Definition, History, and Scholarship." *IEEE Engineering Management Review*. pp.16-31
[6]	M.Mostafa, Mohamed. (2013) "More Than Words: Social Networks' Text Mining For Consumer Brand Sentiments." *Expert Systems with Applications*. pp. 4241-4251
[7]	Go, Alec, Lei Huang, and Richa Bhayani. (2009) "Twitter Sentiment Analysis." Available from: nlp.stanford.edu.
[8]	Ding, Zhikun, Zongjie Li, and Cheng Fan. (2018) "Building Energy Savings: Analysis of Research Trends Based On Text Mining." *Automation in Construction*. pp. 398-410
[9]	Yogi Setiawan, Kadek, Hetti Hidayati, and Alfian Akbar Gozali. (2014) "Analisis User Opinion Twitter Pada Level Fine-grained Sentiment Analysis Terhadap Tokoh Publik [Title in English: *The Analysis of Twetter Opinion User on the Level of Fine-grained Sentiment Analysis toward Public Figures*]." Telkom University, Bandung.
[10]	Liu, Bing. (2012) *Sentiment Analysis and Opinion Mining*, Chicago, Morgan & Claypool.
[11]	Ghiassi, M., and S. Lee. (2018) "A Domain Transferable Lexicon Set for Twitter Sentiment Analysis." *Expert Systems With Applications*. pp. 197-216
[12]	Prayoginingsih, Sila, and Renny Pradina Kusumawardani. (2018) "Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode Support Vector Machine (SVM)." *Jurnal Sisfo*. pp. 83-98.
[13]	Richert, Willi, and Luis Pedro Coelho. (2013) *Building Machine Learning Systems with Python*, in Packt Publishing.
[14]	Ling, Juen, I Putu Eka N. Kencana, and Tjokorda Bagus Oka. (2014) "Analisis Sentimen menggunakan Metode Naïve Bayes Classifier dengan Seleksi Fitur Chi Square [Title in English: *Sentiment Analysis Using Naïve Bayes Classifier Method with Feature Selection of Chi Square*]." *E-Jurnal Matematika*. pp. 92-99.