International Journal of Informatics and Information System ISSN 2579-7069
Vol. 2, No. 1, March 2019, pp. 33-38

Community Opinion Sentiment Analysis on Social Media Using Naive Bayes Algorithm Methods

Taqwa Hariguna [a,*], Vera Rachmawati [a]

[a] Information System Program, STMIK AMIKOM Purwokerto
* corresponding author

Abstract

The election of Governor is an election event for the Regional Head for the future of the region and the country. The Central Java Governor election in 2018 was held jointly on 27 June 2018, which was followed by 2 candidate pairs of the governor. Its many responses from people through twitter's social media to bring up opinions from the public. Sentiment analysis of 2 research objects of Central Java Governor 2018 candidates with a total of 400 tweets with each candidate being 200 tweets. The used of tweets are divided into 3 classes: positive class, neutral class and negative class. In this study the classification process used the Naive Bayes Classifier (NBC) method, while for data preprocessing is using Cleansing, Punctuation Removal, Stopword Removal, and Tokenisation, to determine the sentiment class with the Lexicon Based method produces the highest accuracy in the Ganjar Pranowo dataset with an accuracy of 87,9545%, Precision value is 0.891%, Recall value is 0.88% and F-Measure is 0.851% while Sudirman Said dataset has an accuracy rate of 84.322%, Precision value of 0.867%, Recall value of 0.843% and F-Measure of 0.815%. From these results, we can conclude that the Ganjar Pranowo dataset was higher compared to Sudirman Said's dataset.Click here and insert your abstract text.

Keywords: sentiment analysis, Twitter, Naive Bayes classifier.

1. Introduction

The rapid development of technology and its role in human life can be robed in various activities of human life, either individually or in groups. Approximately terabytes or petabytes of data are contained on the computer network, the World Wide Web (WWW), and the various data storage platforms daily, from business, social, science and engineering, medicine as well as almost every other aspect of daily life [1][2].

Communication 2.0 book reveals that online social media/social network, is not an online mass media because social media has a social power that significantly affects public opinion in the community. Raising support or a mass movement could be due to what is in social media, proven to be able to form opinions, attitudes and public behavior or society [3].

Twitter was a social media created by Jack Dorsey in 2006. In 2013, based on Twitter's Press-release, there were 500 million tweets or Twitter users per day based on the Web Newsportal. A total of 500 million tweets will be free when not utilized while there are a variety of opinion or opinions about the film, celebrities, politicians, products, companies, stocks and events that can be processed into market reference materials or assessments of celebrities, figures, or politicians in the future. Ahead of the elections, political or public figures frequently embolden Twitter to campaign and increase their popularity[4].

The Central Javanese Governor 's election of 2018 was the election of the provincial head of central Java in the period 2018 – 2023. The elections to Central Java Governor year 2018 will be held in unison on 27 June 2018, followed by two prospective couples. Candidate number 1 are H. Ganjar Pranowo, S.H, M.IP and H. Taj Yasin and candidate number 2 are Sudirman Said and Dra. Ida Fauziyah. During the election period, many issues were circulated in the community regarding the Governor of Central Java 2018. The issue of indigenous and expat sentiments can bring down the popularity of other candidates. Issues that develop can interfere with the ideology of pluralism in the country. Prospective candidates should be able to see the potential developed to lift their popularity and their electability [3]. In the context of elections, social media occupies a strategic position as one of the media's campaigns and politic communications. Strategic role of social media in the context of elections that reach the level of popularity and

acceptability [5]. Social Media, especially Twitter, is now one of the most effective and efficient promotional or campaign venues [6].

From the million twit data on Twitter, sentiment analysis can be performed to determine what percentage of positive sentiment is and what percentage of sentiment is negative towards a person, the company, institution, group, or a particular situation [7]. Sentiment Analysis is the detection of attitudes against an object or person [7]. Sentiment analysis refers to the broad field of natural language processing, linguistic computing and text mining.

Application of methods in text mining to classify sentiment into two classes, namely: positive and negative classes such as Support Vector Machine algorithm [8], Naive Bayes [9], Markov blankets and Metaheuristic search [10] and KNNR method [11]. Many works of literature mention that to classify sentiment into two classes, positive and negative, SVM is more accurate than naive Bayes. However, SVM will experience many problems in terms of speed-accuracy to classify sentiment into more than two classes, especially if the classes are not balanced [12].

2. Literature Review

In general, text mining is used to demonstrate a system that analyzes large quantities of quantity from natural language text and detects lexical or linguistic pattern use in order to extracts useful information.

The Naive Bayes algorithm is a simple probability-based predictive technique that is based on the application of the Bayes rules under the assumption of a strong or independent. Besides, Naive Bayes can also analyze the variables that influence the most in the form of opportunity. Naive Bayes is an algorithm or method that is most effective and efficient for the design of machine learning and data mining.[13]

Sentiment analysis is an automated process of understanding, extracting, and processing textual data to obtain the sentiment information contained in a sentence of opinion. Sentiment Analysis is done to see opinions or tendencies toward an issue or object by a person, whether or not it is likely to be a negative or positive opinion.[14]

Twitter is a website owned and operated by Twitter Inc., which offers the social ring Jejaa microblog that allows its users to send and read messages called Twitter (tweet). Twitter (tweet) is a text written that has 140 characters displayed on the user 's profile page. Tweet can widely-viewed, but the sender can restrict the delivery of messages to their friend's list only. Users can see the Twitter author 's tweet with the name of the follower. All users can send and receive tweets on Twitter sites, compatible external applications (mobile phones), or with a short message (SMS) available specific countries. Twitter users, based on PT Bakrie Telecom's data, have 19.5 million users in Indonesia of a total of 500 million users worldwide. Twitter became one of the largest social networks in the world, making it profitable to reach 145 million USD[15].

3. Research Method

3.1. Research Data

The research materials used are public opinion data on Twitter based on the post - comment 2 account of prospective spouse Gurbenur Central Java 2018 verified by Twitter. The account Dataset used in this study was the official n iwas taken : @Ganjarpranowo (ganjar Pranowo) and @sudirmansaid (Sudirman Said). The datasets used are then grouped into three classes, i.e. positive sentiment classes, negative sentiment classes, and neutral sentiment classes. Each word pThere is a data given label according to its class. The Dataset used in This study amounted to 400 comments, from each candidate to 200 comments.

Table 1. Examples of community tweets

Date	*Tweet*
29Maret 2018	*@Saridee2 : Yang ikut arisan korupsi akeh tenan*
26Maret 2018	*@SUNU_adijaya : Pak kenpa ya saya tanya sudah dua kali gk pernah kerespon*
18Mei 2018	*@ajja_anti : Maju terus jadi gub jateng jeh pa.suka*
16Mei 2018	*@annas_cim: Pak ganjar,tolong di desa karanglo,kelurahan karang reja,kab kebumen,ada masalah tentang PDAM,ada mesin air nya tp tidak berfungsi,mohon di cek sama pak Nasim warga mengharapkan bantuan bapak ganjar,,mayoritas pendukung pak ganjar,,,mohon solusi nya*

	tks,,,#GANJARYASIN oke
17Mei 2018	*@addi_auto: Pak tolong wilayah wonogiri timur perangkat desanya di cek lagi pelayananya*

3.2. Preprocessing Data

The following stages of preprocessing are performed

a. Cleansing

This stage aims to clear the words of punctuation or other symbols known as noise. Noise is a form of data that will later interfere with the processing of the data. These are HTML characters, emoticons, @username, # (hastag), URLs, and emails.

b. Punctuation Removal

Punctuation Removal aims to remove existing punctuation marks on datasets such as question marks (?), exclamation(!), dots(.), commas (,) and other punctuation mark.

c. Sentence Normalization

Limitations of the character given by Twitter result in Some words intentionally shortened by the user in order to tweet the user's opinion. There is also some the word has the same meaning, so that the word with the same meaning is dragged. Similarly, foreign languages and dialects or the language of the dialect or slank are converted into Bahasa Indonesia. Hence the process of normalization of sentences is done in order to prevent repeating words that have the same meaning.

d. Tokenization

This tokenization Stage performs the word truncation in the text document into a single word piece. The Snippet is called a token or term.

e. Stemming

In the stemming process, it will find the root word and remove the suffix to the word. Stemming aims to reduce variations of words that have the same base word.

f. Stopword Removal

Stopword Removal is a process for eliminating frequently appearing words but has no influence whatsoever in the extraction of sentiment. Frequently used words like question words (what, who, where, why, why, how, when), and, to, if, this, which, and so on.

g. Determine Sentiment

In this process, the Lexicon Based method is used, which is to determine the sentiment of a word opinion, by scoring the polarity of the opinion P. The polarity score of an opinion P would be worth 1 if the Word is a positive opinion, It is worth 0 if The word is neutral, and worth -1 If the Word is negative.

After determining the value of a sentiment word that contains a positive, neutral, and negative word, it then counts the number of words that contain positive, neutral, and negative opinions.

h. Classification

After going through the pre-processing stage, then the DataSet began to be processed using weka application. Weka uses the document type attribute-Relation File Format (Arff) as the input to classify. This step aims to generate a confusion matrix based on the evaluation method 10-fold Cross validation, where the dataset is divided into 10 subsets (9 subeset as training sets and 1 subsets as testing set) by the number of 10 iterations. Testing done with the algorithm method Naive Bayes.

4. Results and Discussion

DataSet collected from The comments on the Second account post candidate Governor of Central Java 2018 on Twitter and has been through the preprocessing stage, hereinafter the DataSet tested using weka with The format. Arff. The following test results are performed.

The results of the algorithm of Naive Bayes Dataset ganjar Pranowo generate accuracy of 87.9545%. The accuracy value is obtained from the calculation result of Precision, Recall, and F-Measure.

Table 2. Of Confusion Dataset Ganjar Pranowo

	Positive	**Neutral**	**Negative**	**FN (False Negative)**	**TP (True Positive)**
Positive	160	97	0	97	160
Neutral	9	1381	0	0	1381
Negative	0	106	7	113	7
FP (False Positive)	9	203	0		

The result of the algorithm of Naive Bayes datasets Sudirman saaid resulted in accuracy of 84,322%. The accuracy value is derived from the result of Precision, Recall and F-Measure calculations.

Tabel 3. Of Confusion Dataset Sudirman Said

	Positive	**Neutral**	**Negative**	**FN (False Negative)**	**TP (True Positive)**
Positive	149	95	0	95	149
Neutral	0	828	1	1	828
Negative	0	89	18	89	18
FP (False Positive)	0	184	1		

Table 4. Comparison Of The Ganjar Pranowo And Sudirman Said Datasets With Naive Bayes Algorithm

Prospective governors	**Results Accuracy**	**Precission**	**Recall**	**F-Measure**	**Duration**
Ganjar Pranowo	87.9545%	0.891	0.88	0.851	0.1 second
Sudirman Said	84.322%	0.867	0.843	0.815	0 second

In table 4 shows the difference in accuracy obtained from the Ganjar Pranowo and Sudirman Said datasets using the Naive Bayes algorithm of 3.6325%. Time used when running on weka application also beicons between a dataset Ganjar Pranowo and Sudirman Said namely the difference of 0.1 second.

5. Conclusion

From the calculations that have been done on both datasets obtained the accuracy result of each dataset of 87.9545%, with precission value 0.891, Recall 0.88 and F-Measure 0.851 on the Ganjar Pranowo DataSet and 84.322% on Sudirman Said DataSet with precission value 0867, Recall 0,843 and F-Measure 0,815. The Naive Bayes classification method used tends to be stable due to the probability of word occurrence in data. Accuracy value is one of the valuation parameters of the method used, accuracy value in the amount of the number of data that successfully classified correctly according to the class the whole amount of data is classified.

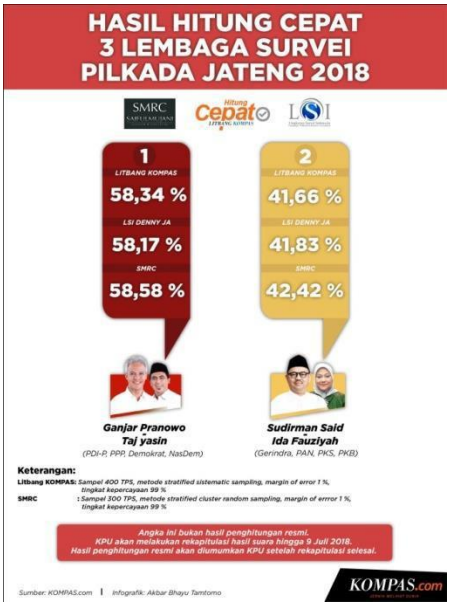

Figure 1. Results of The Survey Institution of The Central Java Elections in 2018

Based on the information on KOMPAS.com shows the final result of quick count conducted three survey institutions for Central Java elections at the 2018 Concurrent elections placing Ganjar Pranowo and Taj Yasin in the first position with 58.34% vote. Meanwhile, Sudirman Said and Ida Fauziyah received a vote of 41.66%.

Thus after seeing the results of the above calculations can conclude that the sentiment of analysis can be withdrawn an Association with the results of the central java elections in 2018.

Analysis of the sentimentof theT Community Opinion on social media Twitter using the Naive Bayes algorithm on This research is still a lot of shortcomings. For further Research to achieve better results, the advice for Subsequent research is expected to further use more data, real Time, and also developed PART Of Speech (POS) Indonesian tagger which is able to improve accuracy in sentiment analysis as well as weighted emoji, using Other algorithms such as CART, C4. 5, SVM AND KKN by comparing the higher level of accuracy.

References

[1] J. Han, and M. Kamber, "Data Mining Concepts And Techniques". Verlag Berlin Heidelberg : Spinger, 2006.

[2] Kompas.com. https://regional.kompas.com/read/2018/03/13/09512101/survei-kompas-banyak-kemungkinan-yang-bisa-terjadi-pada-pilkada-jateng. Accessed 29 September 2018.

[3] A.P. Jain, and V.D. Katkar, "Sentimen analysis of Twitter data using data mining". In 2015 International Conferene on Information Processing (ICIP), 807-810, 2015.

[4] P. Beineke, T. Hastie, C. Manning, and S. Vaithyananthan, "Exploring Sentiment Summarization". In Y. Qu, J. Shanahan, & J. Weibe (eds) Proceedings of the {AAAI} Spring Symposium on Exploring Attitude and Affect in Text : Theories and Applications, AAAI Press, 2004.

[5] Po-Wei Liang, Bi-Ru Dai. Opinion Mining on Social Media Data. IEEE 14th International Conference on Mobile Data Management, pp. 91–96, 2013.

[6] R. Li, K. H. Lei,R. Khadiwala, Chang. TEDAS: A Twitter-based Event Detection and Analysis System. ICDE, pp.1273-1276, 2012 IEEE 28th International Conference on Data Engineering, 2012.

[7] Gonzalez-Marron D., Mejia-Guzman D., Enciso-Gonzalez A. (2017) Exploiting Data of the Twitter Social Network Using Sentiment Analysis. In: Sucar E., Mayora O., Munoz de Cote E. (eds) Applications for Future Internet. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 179. Springer, Cham

[8] Birjali M., Beni-Hssane A., Erritali M. (2017) A Method Proposed for Estimating Depressed Feeling Tendencies of Social Media Users Utilizing Their Data. In: Abraham A., Haqiq A., Alimi A., Mezzour G., Rokbani N., Muda A. (eds) Proceedings of the 16th International Conference on

[9] Hybrid Intelligent Systems (HIS 2016). HIS 2016. Advances in Intelligent Systems and Computing, vol 552. Springer, Cham

[10] Kasturi D. V., Nurhafizah T. Suicide detection system based on Twitter. Science and Information Conference 2014, pp. 785-788, August 27-29, London, UK

[11] Zhao, D. & Rosson, M.B., (n.d), Retrieved from http://research .ihost.com/cscw08-socialnetworkinginorgs/papers/zhao_cscw08_workshop.pdf

[12] Dey P., Sinha A., Roy S. (2015) Social Network Analysis of Different Parameters Derived from Realtime Profile. In: Natarajan R., Barua G., Patra M.R. (eds) Distributed Computing and Internet Technology. ICDCIT 2015. Lecture Notes in Computer Science, vol 8956. Springer, Cham

[13] Yang D., Zheng H., Yan J., Jin Y. (2012) Semantic Social Network Analysis with Text Corpora. In: Tan PN., Chawla S., Ho C.K., Bailey J. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2012. Lecture Notes in Computer Science, vol 7301. Springer, Berlin, Heidelberg

[14] Rawashdeh A., Rawashdeh M., Díaz I., Ralescu A. (2014) Measures of Semantic Similarity of Nodes in a Social Network. In: Laurent A., Strauss O., Bouchon-Meunier B., Yager R.R. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2014. Communications in Computer and Information Science, vol 443. Springer, Cham

[15] James W. Pennebaker, et al. The Development and Psychometric Properties of LIWC2007. The University of Texas at Austin and the University of Auckland, New Zealand, http://www.liwc.net/LIWC2007LanguageManual.pdf