Analysis sentiment in social media against election using the method naive Bayes

Damar Nurcahyono¹, Willy Permana Putra², Abdul Najib³, Tien Rahayu Tulili⁴

¹³⁴ Department of Informatics Engineering, Politeknik Negeri Samarinda, Indonesia
²Department of Informatics, Politeknik Negeri Indramayu

email: ¹damarn@polnes.ac.id; ²willy_p@polindra.ac.id; ³abdulnajib@polnes.ac.id; ⁴tien.tulili@polnes.ac.id

Abstract. Election is the process of selecting people to fill a variety of political positions that are diverse. General elections in a country are usually held periodically where the country adheres to a democratic system. Indonesia is a country that adopts a democratic system. The development of the use of Social Media is very fast. Nowadays, there are many social media which have a big influence. As a result, society has experienced changes in culture, ethics, norms and a more critical mindset in responding to existing conditions as well. Social media is now increasingly easy to use by all groups, from the beginning only a small person to a successful person and famous for Social Media. Sentiment analysis is a science that is useful for analyzing someone's opinions, sentiments, and emotions expressed in the text. In this study the Naïve Bayes method can be applied to the classification of positive, negative, and neutral opinions. In this study of 50 training data and 10 test data obtained an accuracy of 90%. And the results obtained positive sentiment percentage of 50%, negative 20%, neutral 10%. More positive sentiments gained.

1. Introduction

The development of the use of Social Media is very fast. Nowadays, there are many social media which have a big influence. As a result, society has experienced changes in culture, ethics, norms and a more critical mindset in responding to existing conditions as well. Social media is now increasingly easy to be used by people coming from different kind of groups in social, politic or entertainment. Social media is an online media, with its users can easily follow, share, and create content including photos, videos, blogs, social networks and the virtual world. Blogs and social networking are the most common forms of social media used by people throughout Indonesia. Social Media referred to in this research is Twitter. Twitter is a social network that has features that allow users to create and read statuses. A maximum of 140 characters can be uploaded. Users can also see the status of others by becoming a follower of the account of the person they want to see. Status can be used by users to tell about what is being done or felt, conversations, information sharing, and news reporting [1]. In general, tweets are used to post things about users and share information. The contents of the tweet can also express the user's feelings or mood. Some statuses use the symbol ‘#’ which serves as the topic of the discussion to be discussed and may become a trending topic.

Sentiment analysis is a science that is useful for analyzing someone's opinions, sentiments, and emotions expressed in the text. Its function is to classify texts that exist in opinions and sentences. Sentiment Analysis is also called mining opinion [2]. By using text mining to classify opinions of
negative or positive comments on social media comments. This research aims to analyze the comments from social media whether each comment is classified to a positive or negative comment.

2. Literature Review

Naive Bayes is a classification method using probability and statistic method which have been developed by Thomas Bayes. This method predicts future probability based on previous experience and then well known as Bayes Theorem. The main characteristic of Naive Bayes Classifier is a naive assumption about each conditions or actions [3-7].

Advantages of the method is small training data can estimate and classify the data. There are several steps of Naive Bayes, namely [3]: calculate the class/label; calculate the probability for each class; multiply all variable for each class; compare each result for each class.

Here is algorithm formula of Naive Bayes [8, 10].

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

(1)

The explanation for equation 1 as follows:

- **X**: unknown variables data
- **Y**: hypotheses of the data class
- **P(Y|X)**: probability of hypotheses Y according to X condition (posterior probability)
- **P(Y)**: probability of hypotheses Y (prior probability)
- **P(X|Y)**: probability X according to condition in hypotheses Y
- **P(X)**: probability X

One of reliable parameters for a classification algorithm is a prediction and calculated by using accuracy formula.

3. Methodology

The flow diagram of the research concept framework is shown in the following figure 1.

![Figure 1. Research Conceptual Framework](image)

This research was conducted to classify sentiments contained in tweets posted by Twitter users related to Presidential Candidates and Vice-Presidential Candidates in 2019. Starting from data collection and data retrieval, also known as Twitter data crawling using R studio by utilizing Twitter API (Application Programming Interface), which is then saved to the database in the form of MS Excel. Where the data obtained will be classified into several types of sentiment categories, namely negative and positive. The method used to classify is the Naive Bayes Classifier method. Before doing the classification, preprocessing data is needed to simplify the data.
Data Crawling is a technique of taking data automatically from a data storage. In this study the object of the data crawling stage is the Twitter web page. Data obtained from this stage in the form of users and tweets. Furthermore, by using the Twitter API that has been implemented in the PHP programming language, tweet data will be obtained by downloading automatically from the user data obtained.

3.1. Case Folding
From the data that has been in the process all characters with big letters become lowercase

3.2. Filtering
This stage will be cleared of tweets from special characters, URL links, usernames, and emoticons.

3.3. Tokenizing
After passing through the Case Folding stage, the Tokenizing stage continues this stage is the process of breaking down from sentences into words and turning into pieces like words based on each word that makes them up.

3.4. Classification
After pre-processing the words they will be classified into two categories, namely positive and negative opinions.

3.5. Results Preferences
Each candidate will be calculated as a preference value via a tweet collected using the Naive Bayes Classifier.

3.6. Hypothesis
From the framework of the research concepts that have been described, the hypothesis of the research is to apply the Naive Bayes Classifier method to be able to classify positive and negative opinions on 2019 General Election.

4. Experimental Result

4.1. Data Retrieval
The data used are tweets taken on Social Media (Twitter) with the hashtags trending the topic of the Election of Presidential and Vice President Candidates in 2019. Tweet data used are only 50 tweets. Opinion of tweets that have been collected will be through pre-processing. Testing data on the research can be seen in table 4.1
Table 1. Test tweet data (Training Data).

| No | Date       | User         | Tweet                                                                                                                                 |
|----|------------|--------------|---------------------------------------------------------------------------------------------------------------------------------------|
| 1  | 25-Apr-19  | @Dian3w1     | Serba-serbi Pilpres 2019, penuh dengan kecurangan, dan pihak yang dirugikan adalah masyarakat (Miscellaneous of President Election 2019, full of cheating, and people are the victims) |
| 2  | 17-Feb-19  | @YHaryanda   | Siap penuh mendukung pilpres tahun ini. (Ready to support the president election this year)                                           |
|    |            |              | Kami senang dan kami suka jika PBB deklarasi dukung pemilu tahun ini. Insya Alloh akan sangat berkah. (We’re glad and like if the United Nations declared to support the election this year. Hopely, this will be a blessing) |
| 3  | 23-Nov-18  | Sparta5C@    | Melihat perhitungan suara di “beberapa” TPS yang ada di Palembang, Alhamdulillah pilpres tahun ini berjalan lancar. (See the voting count in “several” voting places in Palembang. Thanks God, this election run smoothly) |
| 4  | 20-Apr-19  | @KHMarufAmin_ | begitu juga aku harap para pendukung 02 dan 01, kita berkompetisi dengan cara yang elegan, sportif, jangan fitnah & saling merendahkan. Hargai #pilpres2019 ini sebagai Pesta Demokrasi kita sebagai WNI (So do I. I wish that the supporter of 02 and 01, we compete with an elegant and sportive way, don’t slander and degrade each other. Respect #thepresidentelection2019) |

Table 2. Test tweet data (Training Data).

| No | Tweet                                                                                                                                 |
|----|---------------------------------------------------------------------------------------------------------------------------------------|
| 1  | Sudah banyak masyarakat yang dirugikan akiabat pilres tahun ini. Berdoa saja agar para calon pemimpin bisa mensejahterahkan masyarakat nanti. #pilpres2019 (Many people had got the negative impact because of the election this year. Let’s pray that the candidates can lead the people to prosperity in the future. #thepresidentelection2019) |
| 2  | Jadi pemimpin jangan munafik. #pilpres2019 (do be a goody leader) Yaa kalo kerja yang betul, baru menjadi panitia pemilu saja sudah tidak bisa dipercaya. #pilpres2019 (Please take your work seriously, you are just the election committee but you cannot be trusted. #thepresidentelection2019) |
| 3  | Siapapun yang terpilih di #pilpres2019, adalah pilihan rakyat. #salamdamaial #pilpres2019 (Whoever the elected president in #thepresidentelection2019, he is the choice of the people. #peace) |
| 4  | Terlalu banyak tindakan intimidasi #pilpres2019 (Too much intimidation. #thepresidentelection2019)                                   |
| 5  | Sudah banyak masyarakat yang dirugikan akibat pilres tahun ini. Berdoa saja agar para calon pemimpin bisa mensejahterahkan masyarakat nanti. #pilpres2019 (Many people had got the negative impact because of the election this year. Let’s pray that the candidates can lead the people to prosperity in the future. #thepresidentelection2019) |

4.2. Pre-Processing Stage
Pre-processing is a process of changing the form of data that has not been structured into structured data as needed, for further mining processes.

4.3. Filtering
The first stage is filtering. This stage takes important words from the token results. At this stage the tweet will be cleared of special characters, URL links, usernames, and emoticons. The results of the filtering process can be seen in table 3. The English version is written in the parantheses.
Table 3. Filtering.

| No | Tweet |
|----|-------|
| 1  | Serba serbi Pilpres 2019 penuh dengan kecurangan dan pihak yang dirugikan adalah masyarakat (Miscellaneous of President Election 2019 full of cheating and people are the victims) |
| 2  | Siap penuh mendukung pilres tahun ini (Ready to support the president election this year) Kami senang dan kami suka jika PBB deklarasi dukung pemilu tahun ini Insya Allah akan sangat berkah (We re glad and like if the United Nations declared to support the election this year. Hopely, this will be a blessing) |
| 3  | Mari menuju perubahan yang lebih baik (Let s we go to the better changing) … … begitu juga aku harap para pendukung 02 dan 01 kita berkompetisi dengan cara yang elegan sportif jangan fitnah dan saling merendahkan Hargai pilpres2019 ini sebagai Pesta Demokrasi kita sebagai WNI (So do I. I wish that the supporter of 02 and 01, we compete with an elegant and sportive way, don’t slander and degrade each other. Respect the president election 2019 as our democracy party) |

4.3.1. Tokenizing
The second stage is the tokenizing level. At this stage the words are broken down into words so that they become part of each other's words.

4.3.2. Case folding
The third stage is the case folding stage. Case folding is changing all uppercase or capital letters in Twitter into lowercase letters. Only letters 'a' to 'z' are accepted. Characters other than letters are omitted and are considered delimiter. Delimiter is a sequence of one or more characters used to limit or separate data that is presented in plain text.

4.3.3. Determination of Sentiment
The data obtained will be categorized into three categories: positive, negative and neutral. The determination will be made based on the contents of the tweet data manually. Determination of sentiments can be seen in the table 4.

Table 4. Sentimental Data Dictionary.

| Positive | Negative |
|----------|----------|
| Sejahtera (prosperity) | Damai (peace) | Tertib (orderly) | Dirugikan (be aggrieved) |
| (Dukung (support) | Tenang (quiet) | Berkualitas (good quality) | Bukan (not) |
| Kemaslahatan (benefit) | Kesuksesan (success) | tanggung jawab (responsible) | tidak sesuai (not appropriate) |
| Lancer (smoothly) | Bijak (wise) | Rukun (peace) | Kesalahan (mistake) |
| lebih baik (better) | Maju (go forward) | Jaga (keep) | Intimidasi (intimidation) |
| Jujur (honest) | Kebenaran (truth) | Sportif (sportive) |
| Kemenangan (victory) | Kagum (admire) | Gembira (happy) |
| Jaga (watch) | terima kasih (thank you) | Semangat (spirit) |
| Puji (praise) | Perkembangan (progress) | Lancar (smoothly) |
| Kemenangan (victory) | sukses (success) |

4.4. Naive Bayes Classifiers

4.4.1. Calculates probability
Calculates the i-class prior probability P (Ci) / calculate the number of classes. The results of calculating the number of classes can be seen in table 5.
Table 5. Probability.

| Sentiment | Data |
|-----------|------|
| **Positive** | 39   |
| **Negative** | 6    |
| **Neutral** | 5    |

Table 6. Count the number of classes.

| P(Ci) | P (class sentiment) = “POSITIF” = 39/50 = 0.78 |
|-------|-----------------------------------------------|
|       | P (class sentiment) = “NEGATIF” = 6/50 = 0.12 |
|       | P (class sentiment) = “NETRAL” = 5/50 = 0.10  |

After performing manual calculations with Excel, the next step is to test with the Rapid Miner application to be able to see the accuracy obtained from using the application.

![Naive Bayes Process](image)

**Figure 3. Naive Bayes Process**

![Confusion matrix](image)

**Figure 4. Confusion matrix**

Based on testing using the Rapid Miner application in Figure 4 with the results of calculations done manually, the same results were obtained, namely 90%.

5. Conclusion and Further Work

Based on the results of research that has been done, it was concluded that The Naïve Bayes method can be applied to the classification of positive, negative, and neutral from 50 training data and 10 test data obtained an accuracy of 90%. In the 2019 General Election day on Twitter the percentage of positive sentiment was 50%, negative 30%, neutral 10%. More positive sentiments are got it. For phrases that have positive and negative sentiment cannot classified by NBC because they have different sentiment opinions.
References

[1] Ginting, H. S., Lhaksmana, K. M., & Murdiansyah, D. T. (2018). Klasifikasi Sentimen Terhadap Bakal Calon Gubernur Jawa Barat 2018 di Twitter Menggunakan Naive Bayes. ISSN : 2355-9365 e-Proceeding of Engineering : Vol.5, No.1 Maret 2018, Universitas Telkom, 5(1), 1793–1802.

[2] Zulfa, I., & Winarko, E. (2017). Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 11(2), 187. https://doi.org/10.22146/ijecs.24716

[3] Valle, M. A., Varas, S., & Ruz, G. A. (2012). Job performance prediction in a call center using a naive Bayes classifier. Expert Systems with Applications, 39(11), 9939-9945.

[4] Adityawan, E. (2014). Analisis Sentimen Dengan Klasifikasi Naive Bayes Pada Pesan Twitter Menggunakan Data Seimbang.

[5] Gusriani, S. (2016). ANALISIS SENTIMEN DAN KLASIFIKASI KATEGORI TERHADAP TOKOH PUBLIK PADA TWITTER.

[6] Fathan Hidayatullah, A., Sn, A., Teknik, J., Fakultas, I., & Industri, T. (2014). ISSN: 1979-2328 UPN "Veteran. Seminar Nasional Informatika, 2014(semnasIF), 115–122. Retrieved from http://www.situs.com

[7] Nurhuda, F., & Sihwii, S. W. (2014). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. ITSmart: Jurnal Ilmuhan Teknologi Dan Informasi, 2(2), 35–42.

[8] Riziqiyah, P. (n.d.). ANALISA_SENTIMEN_MENGGUNAKAN_NAIVE_BAYES.

[9] Nurirwan Saputra, Teguh Bharata Adji, A. E. P. (2015). Jurnal Dinamika Informatika Volume 5, Nomor 1, November 2015. Jurnal Dinamika Informatika, 5(November), 12. Retrieved from http://ojs.upy.ac.id/ojs/index.php/dinf/article/view/113

[10] Xue, J., Liu, K., Lu, Z., & Lu, H. (2019, August). Analysis of Chinese Comments on Douban Based on Naive Bayes. In Proceedings of the 2nd International Conference on Big Data Technologies (pp. 121-124).

[11] Ahammed, M. S., Newaz, M. N., & Dey, A. (2019). Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches. Global Journal of Computer Science and Technology.

[12] Silva, G., Costa, M., Drummond, A., & Weigang, L. (2019, July). Predictive Model for Brazilian Presidential Election Based on Analysis of Social Media. In The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (pp. 46-53). Springer, Cham.

[13] Pratama, Y., Tampubolon, A. R., Sianturi, L. D., Manalu, R. D., & Pangaribuan, D. F. (2019, March). Implementation of Sentiment Analysis on Twitter Using Naive Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election. In Journal of Physics: Conference Series (Vol. 1175, No. 1, p. 012102). IOP Publishing.

[14] Habibi, M. N. (2019). Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis.

[15] Hulliyah, K., Bakar, N. S. A. A., Ismail, A. R., & Pratama, M. O. (2019, November). A Benchmark of Modeling for Sentiment Analysis of The Indonesian Presidential Election in 2019. In 2019 7th International Conference on Cyber and IT Service Management (CITSM) (Vol. 7, pp. 1-4). IEEE.

[16] Haryanto, B., Ruldeviyani, Y., Rohman, F., TN, J. D., Magdalena, R., & Muhamad, Y. F. (2019). Facebook Analysis of Community Sentiment on 2019 Indonesian Presidential Candidates from Facebook Opinion Data. Procedia Computer Science, 161, 715-722.

[17] Miranda, E., Aryuni, M., Hariyanto, R., & Surya, E. S. (2019, August). Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content). In 2019 International Conference on Information Management and Technology (ICIMTech) (Vol. 1, pp. 62-67). IEEE.
[18] Oyebode, O., & Orji, R. (2019, October). Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0140-0146). IEEE.

Acknowledgments
This research is fully supported by the Center for Research and Community Service, Samarinda State Polytechnic.