The Implementation of Naïve Bayes Algorithm for Classifying Tweets Containing Hate Speech with Political Motive

[1]R. Reza El Akbar, [2]Rahmi Nur Shofa, [3]Muhammad Ilham Paripurna
Department of Informatics
Siliwangi University
Tasikmalaya, Indonesia
reza@unsil.ac.id, rahmi.shofa@unsil.ac.id,
157006074@student.unsil.ac.id

[4]Supratman
Department of Mathematics Education
Siliwangi University
Tasikmalaya, Indonesia
supratman@unsil.ac.id

***Abstract*— The mid-of 2018 until the mid-of 2019 has been densely marked with political agendas in Indonesia. This moment has created vulnerability to spread hate speech with political motive which is one of the most commonly encountered as cyber-crime on Indonesia's social media. Twitter, as one of the most popular social medias in Indonesia becomes the target of spreading hate speech. Tweets that are suspected to contain hate speech can be drawn automatically using twitter scraper. Filtering and labeling stage before being classified using Naïve Bayes Algorithm. The classification process is done by using WEKA, so finally the accuracy of Naïve Bayes Algorithm for tweet containing hate speech with political motive can be identified. By using that method, the classification can be started and the particular tweets that include non-political hate speech, hate speech with political motive, or non-hate speech can be identified. Data tweet that has been drawn, should then be processed through the average value from the accuracy is 93.4%.**

Keywords*— *Hate Speech with Political Motive, Naïve Bayes, Twitter scraper

I. INTRODUCTION

The rapid development of the internet makes information sources easy to know, this raises the level of vulnerability to the validity of an article [1]. Today, social media is particularly in the spotlight because it allows users to make efforts for manipulating public opinion [2]. In addition to that, the absence of strict regulations makes social media a tool to spread verbal hate speech in the form of certain terms, which has negative connotation [3]. Hate speech is a language used to express hatred towards groups that are targeted or intended to denounce, humiliate, or to insult certain group members [4]. The years of 2018 and 2019 are considered political years in Indonesia, where news will be adorned with political news and co-optation of interests, so it will allow the spread of hate speech on social media.

Based on data revealed by KASATGAS Nusantara at the General Election Discussion entitled Hoax and Law Enforcement, it showed that from mid-2017 to December 2018 there were 3,884 hoax contents and hate speeches. The National Police, detected at least 3,000 accounts that actively spread the utterances of hatred on social media, and 122 of them already arrested.

The survey conducted by *Hootsuite - we are social* in January 2019, showed that Twitter was in the fourth ranks on the most used social network category in Indonesia, which is 52% of the total number of internet users. This makes Twitter a social media that is vulnerable to the rampant speech of politically motivated hatred that has the potentiality to cause riots or divisions in the community.

Free distribution of content on social media causes rampant hoax and information to circulate. The amount of information that circulates quickly makes it difficult to judge the truth of the news that is spread [5]. The rapid development of social media causes a degree of vulnerability to hate speech motivated by religion, politics, social, economic and SARA (*i.e.*, ethnicity, religion, race, and inter-group relation) [6]. In this research focus to classifying on hate speech which political motive.

Tweets from various Twitter users can be classified into several categories which include hate speech with political motive, non-political motives of hate speech and non-hate speech. Data tweets are taken by using *Twitter scraper* so the data can be filtered, labeled and classified using Naïve Bayes Algorithm on the WEKA. Naïve Bayes was chosen because this algorithm can classify documents with their simplicity and speed of computation, but have high competence and good performance on the classification of document data using text numbers [7].

II. THEORETICAL BASIS

A. Hate Speech with Political Motive

Hate speech is a language used to express hatred towards groups that are targeted or intended to denounce, humiliate, or to insult certain group members. Various motives such as religious, political, social and economic motives as well as SARA can be the trigger for the emergence of hate speech that can cause potentials leading to riots [6], [8]. Theoretically, Hate Speech is "a crime motivated by malice or ill will towards a social group" [9].

B. Twitter Scraper

Twitter scraper is a package from Python that contains simple algorithms for pulling out data on Twitter. Data on Twitter is pulled using a command prompt in which

978-1-7281-5192-2/19/$31.00 ©2019 IEEE

command contains keywords, the number of tweets to be taken, the format of the output file, the start date and the end date of the tweet you want to withdraw.

C. Machine Learning

Machine learning is a series of techniques that can help in handling and predicting very large data by presenting these data with learning algorithms [10].

D. Data Mining

Data mining is a process that employs one or more computer learning techniques (machine learning) to analyze and extract knowledge automatically [11].

E. Naïve Bayes Algorithm

Bayes is a simple probabilistic based prediction technique that is based on the application of the Bayes theorem or Bayes rule with the assumption of a strong (naïve) level of independence. The meaning of strong independence is that a feature in a data is not related to the presence or absence of other features in the same data [12].

Below is the general form of the Bayes theorem:

$$P(HIX)P(H) = \frac{P(XIH)P(H)}{P(X)} \quad (1)$$

Where X is data with an unknown class, H is hypothesis of the data X. P(H|X) is probability of hypothesis H based on condition X. P(H) is probability of hypothesis H. P(X|H) is probability X based on the condition, and P(X) is probability of X.

Accuracy values can be calculated using the following formula:

$$Accuracy = \frac{Amount\ Correctly\ Classified}{Total\ Number} x\ 100\% \quad (2)$$

III. Research Methods

The method of this research is divided in three stage and show in Fig. 1.

Fig. 1. Stages of research

A. Data Collection

Data supporting the research was obtained through literature studies and field observations with direct observation techniques, which can be seen in Fig. 2.

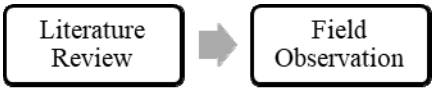

Fig. 2. Data collection process

The literature study used to support this research was a journal related to data mining and machine learning that is used to classify data or documents. Observation was done by direct observation techniques. The process of observation was observing Twitter user submissions in detail and depth.

B. Analysis of the Problem

The problems found in the process of literature study and field observations wreathen examined and found solutions based on the development of existing science and technology. The problem found was the rise of politically motivated hate speech on Twitter.

C. Implementing of the solution

Implementing of the solution divided in three stage, there are crawling, filter & labeling, classifying can be seen in Fig. 3.

Fig. 3. Implementing of the solution's stage

The initial stage of the solution implementation process is 'crawling'. Crawling is the process of retrieving or retrieving data in this study by using twitters to automatically take Twitter user tweets and will be saved in CSV form. The second stage is 'Filter & Labeling'. The filtration process is carried out on a dataset that was previously automatically stored by filtering attributes that only contain the text of the tweet so that it can be labeled whether the tweets include utterances of non-political hatred, speeches of hate politically motivated or non-hate speech. The last stage is the 'Classifying' stage, which is the process of classifying the three categories using the Naïve Bayes algorithm on WEKA tools. The specified stages of Implementing of the solution are in the Fig. 4.

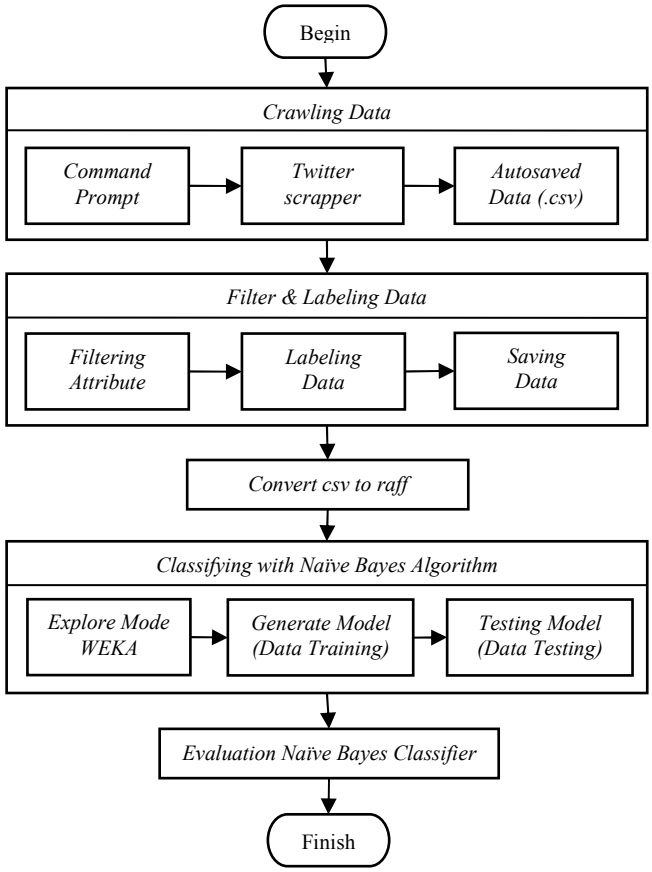

Fig. 4. The flowchart of implementing the solution

1. *Crawling*

The process of data retrieval is done using a specific syntax in the Python program that is accessed via the Command Prompt. The following is the syntax for retrieving tweet data with the keyword "jokodok OR prabohong" whose tweet was sent for the period of 10 April 2019 to 30 May 2019 and then saved to the file "jokodok-prabohong.csv" with a maximum number of strings of 20,000. Below is the following of the complete syntax:

twitterscraper "jokodok OR prabohong" --output jokodok-prabohong.csv –limit 20000 –-begindate 2019-04-10 –enddate 2019-05-30 –csv

The number of data tweets that were successfully drawn in the process were 2,271 tweets and they were automatically saved as *jokodok-prabohong.csv*. The data set was raw data so filtering and labeling must be done so that it can then be classified into three categories namely non-political hate speech, hate speech with political motive and non-hate speech. The initial attributes of the raw data consist of *username, fullname, user_id, tweet_id, tweet_url, timestamp, timestamp_epochs, replies, retweets, likes, is_retweet, retweeter_username, retweeter_userid, retweet_id, text,* and *html*.

2. *Filter and Labeling*

The filtration process removes almost all the attributes that exist in the raw data except the text attribute which will later change its name to the Content attribute. The filtration process is done by importing raw data in CSV format into the Microsoft Excel.xlsx file which is then separated based on the semicolon mark so that it becomes a table. The data of the tweet that has been carried out by filtration is then labeled. So the data attribute consists of *Konten, K1, K2, K3, K4, K5, K6, K7, K8, K9, K10, K11, K12, K13, K14, K15, K16, K17, K18, K19, UK, MF,* dan *Kategori.* Attribute *K1* until *K19* are the keywords that consist of *cebong, ngaciro, mukidi, jokodok, kampret, wowo, praboker, prabohong, jokowi, prabowo, pemilu, presiden, caleg, idiot, goblok, tolol, dungu, bego* and *bajingan*. *UK* is an attribute that indicates if the tweet data contains hate speech. *MF* is an attribute that indicates if the tweet data contains political motive of hate speech, and *Kategori* is an attribute that contains labeling of tweet data included in the *UK-Non-Pol, UK-Pol* or *Non-UK categories*.

Below is the following of Microsoft Excel's formula to indicate and determine which tweet contains political motive and the sample of hate speech that have defined before.

- *=IF(ISNUMBER(SEARCH("cebong",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("ngaciro",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("mukidi",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("jokodok",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("kampret",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("wowo",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("praboker",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("prabohong",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("jokowi",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("prabowo",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("pemilu",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("presiden",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("caleg",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("idiot",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("goblok",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("tolol",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("dungu",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("bego",A2)), "Yes", "No")*
- *=IF(ISNUMBER(SEARCH("bajingan",A2)), "Yes", "No")*

Formula *IF(ISNUMBER(SEARCH))* is used to look for a string that is in a particular cell, in that example cell *A2*. The string enclosed by quotation marks is a sample of hate speech that has previously been defined. If cell *A2* contains words in quotation marks, then in the cell filled in the formula will be filled with the string "*Yes*". If cell *A2* does not contain words in quotation marks, then in the cell filled in the formula will be filled with the string "*No*". The next process is to validate any words that can reinforce a sentence into the hate speech category. The sample words taken are words that are commonly used on social media to express insults against the other person such as *idiots, stupid, and bastards*. Each expression is stored in *O2, P2, Q2, R2, S2* and *T2* cells. Below is the formula used in this process:

=IF(O2="Yes", "True", IF(P2="Yes", "True", IF(Q2="Yes", "True",IF(R2="Yes","True",IF(S2="Yes","True",IF(T2="Yes","True","False")))))

The formula is an IF series formula where if is the condition of one of the cells of *O2, P2, Q2, R2, S2* and *T2* is *Yes*, then the condition generated from that formula will be *True*. If the six cells are *No*, then the condition generated from the formula will be *False*. The result of the formula is entered into cell *U2* whose cell attribute is named *UK*. The next process is to label the tweet data whether it contains political motives or not. Sample words taken are words commonly used on social media to express political sentiment such as *cebong, ngaciro, mukidi, jokodok, kampret, wowo, praboker, prabohong, jokowi, prabowo, pemilu, presiden,* and *caleg*. These expressions are saved in cell *B2, C2, D2, E2, F2, G2, H2, I2, J2, K2, L2, M2* and *N2*. Below is the formula used in this process:

=IF(B2="Yes", "True", IF(C2="Yes", "True", IF(D2="Yes", "True",IF(E2="Yes","True",IF(F2="Yes","True",IF(G2="Yes","True",IF(H2="Yes","True",IF(I2="Yes","True",IF(J2="Yes","True",IF(K2="Yes","True",IF(L2="Yes","True",IF(M2="Yes","True",IF(N2="Yes","True","False")))))))))))))

The formula is an IF series formula where when the condition of one of the cells of *B2, C2, D2, E2, F2, G2, H2, I2, J2, K2, L2, M2* and *N2* is *Yes,* then the conditions resulting from this formula will be *True*. If the thirteen cells are *No*, then the condition generated from the formula will be *False*. The result of the formula is entered in cell *V2* whose cell attribute is named *MF*. The last process is a labeling process whether the data of the previous tweet has been taken including non-political hate speech, hate speech with political motive or non-hate speech. Below is the formula that indicates if the tweet include to the non-political hate speech:

=IF(U2="True", IF(V2="True","UK-Pol","UK-Non-Pol"), IF(U2 ="False",IF(V2="True","Non-UK","Non-UK")))

The formula shows that if the conditions of cell *U2* and *V2* are *True*, then the conclusion will be included in the *UK-Pol* category which is a statement of hate speech with political motive. If the condition of cell *U2* is *True* but *V2* is *False*, then the conclusion will be included in the *UK-Non-Pol* category which is a statement of non-political hate speech. *Non-UK* category (non hate speech) will be obtained if cell *V2* is *True* or *False*, and *U2* is *False*. The following is a condition table that describes the labeling process for tweets whether the category of non-political hate speech, hate speech is politically motivated or non-hate speech. The result of the formula is entered into cell *W2* in which cell attribute is named Category as stated in Table I.

TABLE I. THE CATEGORY OF HATE SPEECH'S CONDITION

U2	V2	Kategori (W2)
True	*True*	*UK-Pol*
True	*False*	*UK-Non-Pol*
False	*True*	*Non-UK*
False	*False*	*Non-UK*

Examples of data on tweets that have passed the Labeling process for speech categories of non-political hate speech, hate speech with political motive and non hate speech are explained in Table II.

TABLE II. THE EXAMPLE OF LABELING TWEET

UK-Non-Pol	UK-Pol	Non-UK
Rizky Rahmat Putra legend tolol bego. Udah setahun main moba gx pernah mythic. Dasar bego.	Rezim ANJING.... BANGSAT KAU JOKODOK.... PKI	Lebih annoying lagi sama beberapa kpopers yang merasa sok superior sama non kpoper.

Data that has passed the filter and labeling process, then was converted from the CSV form to the ARFF form using the supporting tools namely Notepad ++. The purpose of the process is so that the dataset can be accessed or opened using WEKA so that later it can be classified using the Naïve Bayes algorithm.

3. Classifying

Classifying is the last stage of the solution implementation process. The process carried out at this stage is to calculate the classification of datasets that have previously been through the process of Crawling, Filtering and Labeling.

IV. RESULT AND DISCUSSION

The classification process is carried out by using WEKA tools. The following is the calculation of tweet data by category non-political hate speech (UK-Non-Pol), motivated speech politics (UK-Pol), and non-hate speech (Non-UK) described in Fig. 5.

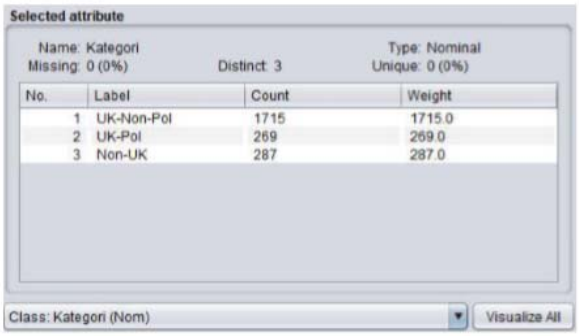

Fig. 5. Attribute calculation based on hate speech categories

The calculation shows that of 2,271 tweets in the dataset, there are as many 1,715 indicated as an expression of non-political hate speech, 269 indicated as an expression of politically motivated hate speech, and 287 is indicated as non hate speech.

The next process is to test between training data with testing data. Testing is done 3 times. The first was conducted on 2,271 training data and 2,000 test. The second testing was conducted on 2,271 training data and 1,500 test data. The third testing is conducted on 2,271 training data and 1,220 test data.

Testing I

Testing I was conducted using 2,271 training data and 2,000 test data. First of all, the Naïve Bayes algorithm is trained first with using training data. After that, the test data is inserted to calculate the accuracy value. Overall the accuracy value obtained is 98.7%

Testing II

Testing II was conducted using 2,271 training data and 1,500 test data. First of all, the Naïve Bayes algorithm is trained first using training data. After that, the test data is inserted to calculate the accuracy value. Overall the accuracy value obtained is 94.4%

Testing III

Testing III was conducted using 2,271 training data and 1,220 test data. First of all, the Naïve Bayes algorithm is trained first using training data. After that, the test data is inserted to calculate the accuracy value. Overall the accuracy value obtained is 86.56%

Below is the summary of overall testing result with accuracy average values can be seen in Table III.

TABLE III. SUMMARY OF OVERALL TESTING

Number of Test	Total Training Data	Total Testing Data	Percentage Accuracy
I	2.271	2.000	98,7%
II	2.271	1.832	94,4%
III	2.271	1.560	86,5574%
Average			93,22%

The absolute error average values from first, second, and third testing can be seen in Table IV.

TABLE IV. ABSOLUTE ERROR AVERAGE VALUES

Number of Test	Error Values	
I	0.0024	0,02%
II	0.0059	0,06%
III	0.0004	0,04%
	Average	0,04%

V. CONCLUSIONS

Based on the experiments that have been carried out, some conclusions can be drawn include the following of:

- The process of withdrawing the tweet data is done using *twitterscraper* via the command prompt. The attributes of the tweet data obtained consist of *username, fullname, user_id, tweet_id, tweet_url, timestamp, timestamp_epochs, replies, retweets, likes, is_retweet, retweeter_username, retweeter_userid, retweet_id, text,* and *html*. The tweet data can be used as a dataset after filtering and labeling so that the attributes become *Konten, K1, K2, K3, K4, K5, K6, K7, K8, K9, K10, K11, K12, K13, K14, K15, K16, K17, K18, K19, UK, MF*, and *Kategori*.
- The classification process of the dataset resulted in three categories of hate speech that consist of non-political hate speech, hate speech with political motive and non-hate speech. The test results of the Naïve Bayes algorithm for the hate speech with political motive carried out on WEKA tools. The average value's result of accuracy is 93.22%

Based on the results of the study, further research might be done with the following parameter changes:

- The time period for taking a tweet data can be defined longer on twitters so that it will produce more datasets.
- The process of labeling tweet data can be done by adding or defining a number of additional keywords as a sample of utterances of hatred and political sentiment, resulting in a more accurate process of categorizing hate speech with political motives.
- The testing process can be done several times with a combination of numbers between training data and different test data in each test, so we can identify how optimal the Naïve Bayes algorithm when used for the classification of politically motivated hate speech.

REFERENCES

[1] D. Maulina, R. Sagara, "Klasifikasi artikel hoax menggunakan support vector machine linear dengan pembobotan term frequency – Inverse document frequency," Jurnal Manajemen Informatika, vol. 2, no. 1, 2018.

[2] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 Russian interference twitter campaign," IEEE/ACM ASONAM International Conference on Advances in Social Networks Analysis and Mining, 2018.

[3] L. Herlina, "Disintegrasi sosial dalam konten media sosial Facebook," Jurnal Pembangunan Sosial, vol. 1, no. 2, 2018, pp. 232-258,.

[4] T. Davidson et al., "Automated hate speech detection and the problem of offensive language," the Eleventh International AAAI Conference on Web and Social Media (ICWSM), 2017.

[5] E. Tacchini, G. Ballarin, M. L. D. Vedova, S. O. Moret, L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," 2018. [Online]. Available: https://arxiv.org/pdf/1704.07506.pdf

[6] M. M. Munir, M. A. Fauzi, R. S. Perdana**, "**Implementasi metode backpropagation neural network berbasis lexicon based features dan bag of words untuk identifikasi ujaran kebencian pada Twitter", Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 10, 2018, pp. 3182-3191.

[7] A. P. Wijaya and H. A. Santoso, **"**Naive Bayes classification pada klasifikasi dokumen untuk identifikasi konten e-government," Journal of Applied Intelligent System, vol. 1, no. 1, 2016, pp. 48-55.

[8] R. Hamad, "Hate crime: Causes, motivations and effective interventions for criminal justice social work," City of Edinburgh Council, 2017. [Online]. Available: https://cycj.org.uk/wp-content/uploads/2017/06/Hate-Crime-causes-and-motivations.pdf

[9] K. P. Danukusumo, "Implementasi deep learning menggunakan convolutional neural network untuk klasifikasi citra candi berbasis Gpu,", S1 Thesis, Universitas Atma Jaya Yogyakarta, 2017.

[10] H. Widayu et al, **"**Data mining untuk memprediksi jenis transaksi nasabah pada koperasi simpan pinjam dengan algoritma *C4.5*," Media Informatika Budidarma, vol. 1, no. 2, 2017, pp. 32-37.

[11] S. Anisah, A. S. Honggowibowo, and A. Pujiastuti, "Klasifikasi teks menggunakan chi square feature selection untuk menentukan komik berdasarkan periode, materi dan fisik dengan algoritma Naivebayes. COMPILER, vol. 5, no. 2, 2016, pp. 59-66.