THE IMPLEMENTATION OF THE MACHINE LEARNING ALGORITHM FOR THE SENTIMENT ANALYSIS OF INDONESIA'S 2019 PRESIDENTIAL ELECTION

by Ghulam Buntoro

Submission date: 03-Jul-2020 10:15AM (UTC+0700)
Submission ID: 1352863325
File name: Ghulam_Jurnal_IIUM.docx (587.96K)
Word count: 5424
Character count: 30925

https://doi.org/10.31436/iiumej.vxxixx.xxxx

THE IMPLEMENTATION OF THE MACHINE LEARNING ALGORITHM FOR THE SENTIMENT ANALYSIS OF INDONESIA'S 2019 PRESIDENTIAL ELECTION

GHULAM ASROFI BUNTORO[1*], RIZAL ARIFIN[1], GUS NANANG SYAIFUDDIIN[1], ALI SELAMAT[2,3], ONDREJ KREJCAR[3], HERRERA-VIEDMA[4], HAMIDO FUJITA[5],

[1]*Faculty of Engineering, Universitas Muhammadiyah Ponorogo, Indonesia*
[2]*Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia*
[3]*Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic*
[4] *Andalusian Research Institute DaSCI "Data Science and Computational Intelligence", University of Granada, Spain*
[5] *Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan*

**Corresponding author: ghulam@umpo.ac.id*

(Received: Day Month Year; Accepted: Day Month Year; Published on-line: Day Month Year)

ABSTRACT: Recently, Indonesian people have participated in the democratic process of electing a new president, vice president, and various legislative candidates for the country. The 2019 Indonesian presidential election was very tense in terms of the candidates' campaigns in cyberspace, especially on social media sites such as Facebook, Twitter, Instagram, Google+, Tumblr, LinkedIn, etc. The Indonesian people used social media platforms to express their positive, neutral, and also negative opinions on the respective presidential candidates. The campaigning of respective social media users on their choice of candidatures for regents, governors, and legislative up to presidential candidates was conducted via the internet and online media. Therefore, the aim of this paper is to conduct sentiment analysis on the candidates in the 2019 Indonesia presidential election based on Twitter datasets. The study used datasets on the opinions expressed by the Indonesian people available on Twitter with the hashtags (#) containing "Jokowi and Prabowo." We conducted data preprocessing using a selection of comments, data cleansing, text parsing, sentence normalization, and tokenization based on the given text in the Indonesian language, determination of class attributes, and, finally, we classified the Twitter posts with the hashtags (#) using Naïve Bayes Classifier (NBC) and a Support Vector Machine (SVM) to achieve an optimal and maximum optimization accuracy. The study provides benefits in terms of helping the community to research opinions on Twitter that contain positive, neutral, or negative sentiments. Sentiment Analysis on the candidates in the 2019 Indonesian presidential election on Twitter using non-conventional processes resulted in cost, time, and effort savings. This research proved that the combination of the SVM machine learning algorithm and alphabetic tokenization produced the highest accuracy value of 79.02%. While the lowest accuracy value in this study was obtained with a combination of the NBC machine learning algorithm and N-gram tokenization with an accuracy value of 44.94%.

ABSTRAK: Baru-baru ini, orang Indonesia telah berpartisipasi dalam proses demokrasi memilih presiden baru, wakil presiden, dan berbagai calon legislatif untuk negara tersebut. Pemilihan presiden Indonesia 2019 sangat tegang dalam hal kempen calon di ruang siber, terutama di laman media sosial seperti Facebook, Twitter, Instagram, Google+, Tumblr, LinkedIn, dll. Orang Indonesia menggunakan platform media sosial untuk menyatakan positif mereka , berkecuali, dan juga pendapat negatif terhadap calon presiden masing-masing. Kampanye pengguna media sosial masing-masing mengenai pencalonan pilihan mereka untuk bupati, gabenor, dan perundangan hingga calon presiden dilakukan melalui media

internet dan dalam talian. Oleh itu, tujuan makalah ini adalah untuk melakukan analisis sentimen terhadap calon dalam pemilihan presiden Indonesia 2019 berdasarkan kumpulan data Twitter. Kajian ini menggunakan kumpulan data mengenai pendapat yang diungkapkan oleh orang Indonesia yang terdapat di Twitter dengan hashtag (#) yang mengandungi "Jokowi dan Prabowo." Kami melakukan pemrosesan data menggunakan pilihan komentar, pembersihan data, penguraian teks, normalisasi kalimat, dan tokenisasi berdasarkan teks yang diberikan dalam bahasa Indonesia, penentuan atribut kelas, dan, akhirnya, kami mengklasifikasikan catatan Twitter dengan hashtag (#) menggunakan Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM) untuk mencapai ketepatan pengoptimuman yang optimum dan maksimum. Kajian ini memberikan faedah dari segi membantu masyarakat meneliti pendapat di Twitter yang mengandungi sentimen positif, neutral, atau negatif. Analisis Sentimen terhadap calon dalam pemilihan presiden Indonesia 2019 di Twitter menggunakan proses bukan konvensional menghasilkan penjimatan kos, waktu, dan usaha. Penyelidikan ini membuktikan bahawa gabungan algoritma pembelajaran mesin SVM dan tokenisasi abjad menghasilkan nilai ketepatan tertinggi iaitu 79.02%. Manakala nilai ketepatan terendah dalam kajian ini diperoleh dengan kombinasi algoritma pembelajaran mesin NBC dan tokenisasi N-gram dengan nilai ketepatan 44.94%.

KEY WORDS: *sentiment analysis, president, Indonesia, naïve Bayes classifier, support vector machine*

1. INTRODUCTION

The turmoil resulting from organizing an Indonesian general election, notably the presidential election of 2019, has been felt since last year. This has applied not only in the real world but also in cyberspace, mainly on social media sites such as Twitter, Instagram, Facebook, etc., which people used to discuss their potential presidential candidates. The stages of the general election and presidential election in 2019 were announced by the Indonesian General Elections Commission (KPU). The names of the presidential candidates had been widely discussed on social media as far back as the candidates' registration phase by the early 2019 Indonesian KPU[1]. The virtual world is a world that is so free and difficult to control, where everyone is free to speak or give their opinion on their respective candidates. The opinions expressed by the public may be positive or neutral but also negative.

The world of information has developed so fast that there is now a significant amount of online media, from news information to social media or friendships, with social media starting from Facebook, Twitter, Path, Instagram, Google+, and many more. Twitter has a total of 330 million active users to date, while around 500 million tweets are made worldwide every day. There are around 100 million active daily users of Twitter around the world[2].

Social media is not only used as a means of friendship or for making friends but also for activities such as the promotion of merchandise or sale and purchase, up to political party promos or campaigns for regent, presidential, and legislative candidates. The team charged with ensuring a candidate for president or regional head, for example, is successful will justify any means of campaigning for their candidate, as evidenced by the presence of many Black Campaigns during the campaign period[3], especially on social media against a candidate. Today's campaign or imaging is not only done in the real world but also in the virtual world. Social media, especially Twitter, is now one of the most effective and efficient campaign venues.

Sentiment analysis continues to be used as part of opinion mining research. It is the process of understanding, extracting, and processing textual data automatically to obtain the sentiment information contained in an opinion sentence[4].

In this study, sentiment analysis was conducted with the aim of viewing and retrieving information pertaining to the opinions expressed by people in the Indonesian language on Twitter with regard to the candidates in the 2019 Indonesian presidential election, whether those opinions were in the category of positive, neutral, or negative. To test the accuracy of the sentiment analysis in this study, we use two machine learning algorithms, namely Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) and 7 tokenizations including an Alphabetic Tokenizer, Character N-gram Tokenizer, Unigram, Bigram, Trigram, N-gram, and Word Tokenizer. The result will enable us to see the accuracy of the machine learning methods NBC and SVM[5] and 7 tokenizations including the Alphabet Tokenizer, Character N-gram Tokenizer, Character Tokenizer, Unigram, Bigram, Trigram, N-gram, and Word Tokenizer for sentiment analysis of the 2019 Indonesian presidential candidates.

2. RELATED WORK

Sentiment analysis research used machine learning to classify Turkish political news[6]. This research classified the sentiment toward Turkish political news and determined whether the sentiment expressed was positive or negative. The different features of Turkish political news were extracted with the machine learning algorithms Naïve Bayes Classifier (NBC), Maximum Entropy (ME), and SVM to produce a classification model.

Sentiment analysis is used to group texts according to their positive or negative orientation[7]. This paper explains the experimental results that apply SVMs to conduct benchmarking with standard datasets to train sentiment analysis classifiers. N-grams and different weighting schemes were used to extract the most classic features. This study also explores the Chi-Square weight feature to select informative features for the classification method. The results of this experimental analysis reveal that using the Chi-Square feature selection can significantly enhance classification accuracy.

The main challenge for law enforcement in recent years has been the automatic detection of abusive language in online media[8]. First, we have developed a deep learning architecture that uses word frequency vectorization to implement the features above. Second, we have proposed a method that, because it does not use pre-trained word embeddings, is an independent language. Third, we have conducted a comprehensive evaluation of our model using public datasets from labeled tweets, open-source implementations built using Keras. The paper presents an ensemble classifier for detecting hate speech in short texts, such as opinion tweets used as corpus datasets[9]. Our classification uses deep learning and combines a set of features related to user behavior characteristics, such as the tendency to send rough messages as input to a combination of machine learning algorithms[10][11].

Sentiment analysis research was carried out by using a hybrid approach[12] with his research methods, including mining association rules, parsing dependencies, and Sentiwordnet applied to solve this aspect-based sentiment analysis problem[13]. The performance of the research was evaluated using negative and racial domains and other benchmarks to evaluate the accuracy of aspect-based sentiment classification.

https://doi.org/10.31436/iiumej.vxxixx.xxxx

3. PROPOSED METHODOLOGY

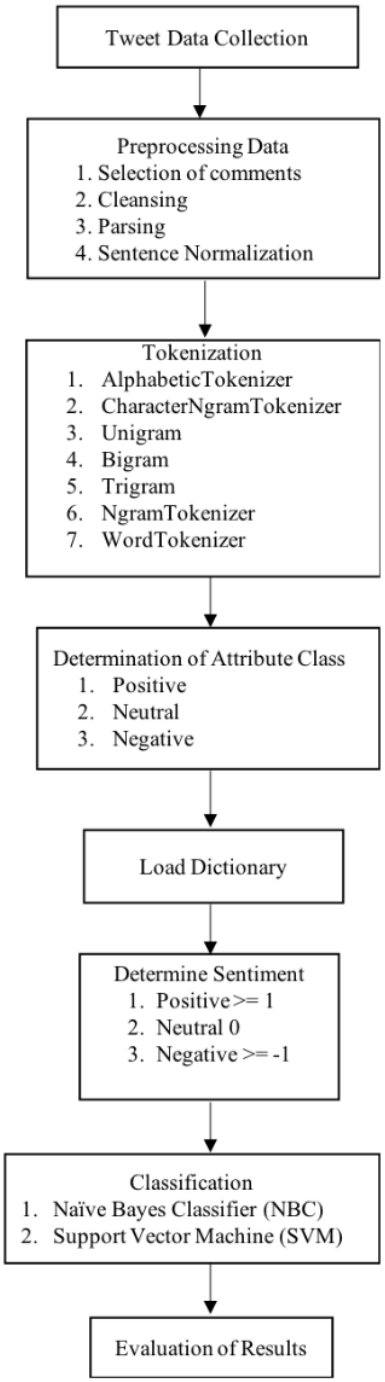

Fig. 1. Proposed Method of Twitter Sentiment Analysis

https://doi.org/10.31436/iiumej.vxxixx.xxxx

3.1.Tweet Data Collection

Crawling[14] carried out tweet data collection with R Programming using R-Studio from Twitter. The data taken comprised only tweets in Indonesian, which consisted of 5000 tweets containing the Jokowi keywords and 5000 tweets containing the Prabowo keywords, to give a total of 10000 tweets. The data were taken randomly from ordinary users of Twitter.

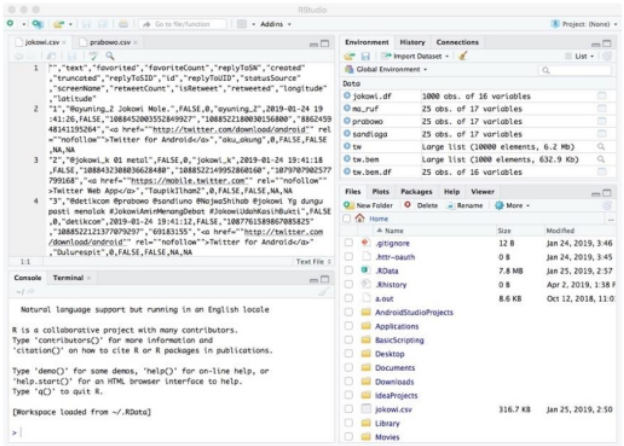

Fig. 2. Crawling tweets of opinions about Jokowi from Twitter at R-Studio

Figure 2 shows the coding of R Language[15] that was used to crawl the data from Twitter. The tweet data on Jokowi comprised 5000 tweets, which, along with examples of original tweets, featured lots of noise characterized by the presence of symbols and links.

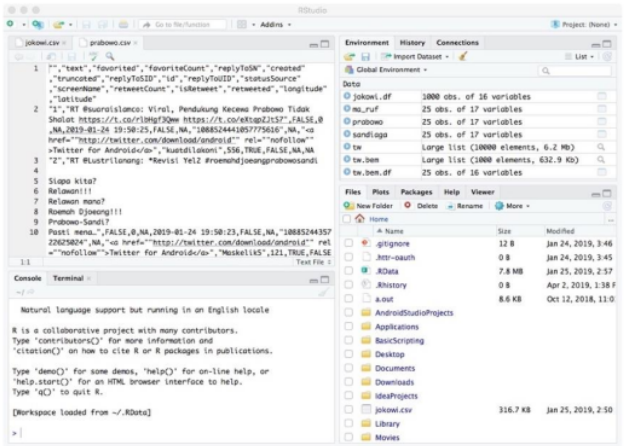

Fig. 3. Crawling tweets of opinion about Prabowo from Twitter at R-Studio

Figure 3 contains a coding of the R Language for crawling data from Twitter. Also seen is the number of data tweets taken about Prabowo, which numbered 5000. The left-hand panes in Figures 2 and 3 also contain examples of tweets in their original form and lots of noise from Twitter with the presence of symbols and links.

3.2. Data preprocessing

The data preprocessing stage[16] in this study consisted of 4 steps, which are described as follows:

1. Selection of Comments

At this stage, comments were selected that contained the keywords of hashtags (#) Jokowi and Prabowo; any data that did not contain both were deleted. When crawling all comments with the hashtag, both will be taken even if they appear in the same sentence. Then, during this process, the same comment will be deleted, even if it comes from a different Twitter account, in order to find unique tweet data.

2. Cleansing

This process aimed to clean up any comments from Twitter that were still dirty and contained a lot of noise. The opinion sentences obtained from Twitter usually contained a certain level of noise, i.e., random errors or variants in measured variables; therefore, we had to eliminate and clean the noise. The items omitted were usually HTML characters, symbols, emoticon icons, hashtags (#), usernames (@username), URL addresses (http://websitename.com), and email addresses (name@websitename.com).

3. Parsing

The third data preprocessing step in this study was parsing[17]. The aim was to break the document into a string of words and then analyze the collection of words by separating them and determining the syntactic structure of each word.

4. Sentence Normalization

The aim of this step was to normalize the sentences taken from Twitter; for example, a sentence containing the words Gaul or Alay[18] would be normalized so that the sentence or language of Gaul and Alay could be recognized as a language following KBBI (The Great Dictionary of the Indonesian Language)[19]. The normalization of sentences involved the following processes:

- Stretch punctuation and symbols other than the alphabet

Stretching punctuation involves inserting distance around the punctuation associated with words that come after or before. The aim is to avoid any punctuation and/or symbols other than those in the alphabet becoming one with the words during the tokenization process.

- Change to all lowercase letters
- Normalization of words

The rules in the normalization process are shown in Table 1.

Table 1: Rules for normalizing words

Normal / slang	**Normal**
Suffix –ny	Suffix –nya
Suffix –nk	Suffix –ng

Suffix –x	Suffix –nya
Suffix –z	Suffix –s
Suffix –dh	Suffix –t
Repeat words: sama2	Repeat words: sama-sama
Spelling: oe	Alphabet: u
Spelling: dj	Alphabet: j

- Eliminate repeated letters

When happy or upset, someone may write opinions based on their emotions; often, when expressing this in written form, they will repeat the same letter. For example: "kereeen" to express pleasure. Repeated words like "kereeen" will be normalized to "cool".

3.3. Tokenization

After normalizing the sentence, it was then broken down into tokens[20] using a delimiter or space bar. The tokens used in this study are:

- Alphabetic Tokenizer: These tokens are formed only from adjacent alphabetical sequences, for example: aku, anak, asli, baik, bagus, cara, cinta, demi, engkau, enak, film
- Character N-gram Tokenizer: This tokenizer divides token is a one-word character; for example: pe, mi, lu, pe, mi, li, han, u, mum
- Unigram: This tokenizer divides the sentence into a token, with each token consisting of only one word; for example, "Pemilu".
- Bigram: This tokenizer divides the sentence into a token, with each token consisting of two words; for example: "Pemilihan Umum.
- Trigram: This tokenizer divides the sentence into a token, with each token consisting of three words; for example, Pemilihan Umum Indonesia".
- N-gram Tokenizer: This tokenizer divides the string into n-grams with the minimum and maximum number of grams as specified; for example, "pemilihan, pemilihan umum, pemilihan umum Indonesia, aku, aku anak, aku anak indonesia"
- Word Tokenizer: This tokenizer divides token from the basic words; for example, "aku, akun, akuntansi, alam, alami, alamiah"

3.4. Determination of Class Attribute

After preprocessing, the next stage in this research is to determine the class attribute. The class attribute used here is sentiment class; in this study, there are 3 class attributes[21], namely positive, neutral, and negative. The use of 3 class attributes provides a more detailed and accurate classification of public opinion toward a particular object.

3.5. Load Dictionary

Following the class attribute determination, the next step is to apply the Lexicon-based method[22]. The dictionary used in this study comprises positive words (positive keywords), negative words (negative keywords), and negation words (negation keywords).

The following is a sample dictionary and its contents:

- Positive keywords; for example, “amanah, ahli, jujur, adil, keren”.
- Negative keywords; for example, “apatis, benci, dosa, jahat, buruk”.
- Negation keywords; for example, “lebih, kurang, tidak, bukan”.
- Dictionary of slang conversion to KBBA; for example, “cyg = sayang, lbh = lebih, krn = karena, jd = jadi, spt = seperti, ciyus = serius”.

3.6. Determination of Sentiment

This is the process used for determining the sentiment (Positive, Neutral, or Negative) in Twitter data once the processing has been performed. The sentiment determining process used in this study consisted of the Lexicon-based or Dictionary-based method with Python Programming. In this study, we are using the Positive and Negative Dictionary. The polarity score of an opinion word (p) will be 1 if the word is in the positive dictionary, meaning the word is positive. A word that is in neither the positive nor negative dictionary is worth 0, meaning it is neutral, while a word in the negative dictionary is worth -1, meaning it is negative[23]. The method for determining sentiment uses the sum formula n, namely the opinion polarity score of the word, plus p, that is the opinion commenting on the feature (f).

After determining which words in a Twitter opinion sentence are positive, neutral, or negative, the weight of the values contained in the sentence is then calculated by totaling the value of each opinion word. If the number of opinion words in the sentence is > = 1, then the sentiment value of the opinion sentence is positive; if the opinion value of the sentence is 0, then the sentiment value of the opinion sentence is neutral, and if the opinion word value in the sentence is > = -1, then the sentiment value of the opinion sentence is negative. The determination of sentiment can be seen in Table 2.

Table 2: Determination of Sentiment

Sentiment	Value
Positive	>=1
Neutral	0
Negative	>=-1

3.7. Classification processes

Following the process for determining sentiment and having established the sentiment value of each opinion sentence using Python Programming, the next step is the sentiment classification process. The classification process uses the WEKA 3.8.3 Machine Learning

https://doi.org/10.31436/iiumej.vxxixx.xxxx

tool[24], and the machine learning algorithms used in this study are NBC and SVM. In the classification process, the data were tested using the 10-fold cross-validation method[25]. The method works by dividing the dataset into two, namely 10 parts with 9/10 parts used as training data and 1/10 parts used as testing data. The iteration process in the method can be run 10 times with a variety of training data and data testing using a combination of 10 parts of data.

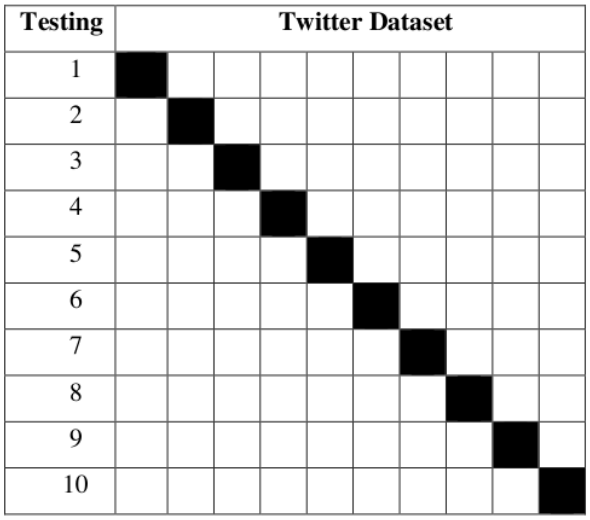

Fig. 4. Illustration of 10-fold cross-validation

3.8. Evaluation of results

The stages of evaluation in the study will examine the performance of Accuracy, Precision, and Recall from the experiments that have been carried out. The results evaluation process is conducted using a Confusion Matrix[26] featuring as its indicators a true positive rate (TP rate), true negative rate (TN rate), false positive rate (FP rate), and false negative rate (FN rate). The TP rate is the percentage of the positive class that is successfully classified as a positive class, while the TN rate is the percentage of the negative class that is successfully classified as a negative class. The FP rate is a negative class that is classified as a positive class, and the FN rate is a positive class that is classified as a negative class.

Table 3: Confusion Matrix

		Predicted		
		Positive (A)	Neutral (B)	Negative (C)
Actual	Positive (A)	AA	AB	AB
	Neutral (B)	BA	BB	BC
	Negative (C)	CA	CB	CC

https://doi.org/10.31436/iiumej.vxxixx.xxxx

4. EXPERIMENT AND RESULTS

In this study, the dataset was derived from tweets of public opinion on the Indonesian 2019 presidential candidates. The data were taken using the crawling method[27] with R Programming using R-Studio from Twitter social media. The data taken were only tweets in Indonesian, with the details of 5000 tweets containing Jokowi's keywords and 5000 tweets containing Prabowo's keywords, giving a total of 10000 tweets. The tweet data were taken randomly from both ordinary users and from the news media online on Twitter.

Following the data preprocessing, tokenization, and class attribute determination steps, the dataset used for this study contained opinion sentences from Twitter classified into their respective sentiment classes (Positive, Neutral, or Negative) with Python Programming. The number of datasets is not the same as the amount of data taken because, during the data preprocessing, the same opinion sentence will be deleted to search for unique data, whereas when the data are being crawled, all opinion sentences will be taken even though the sentence is the same. Table 4 contains the results of the determination of the sentiment class using the Lexicon-based method[28] in Python Programming with three attribute classes, namely positive, neutral, and negative.

Table 4: Results of Determination of Sentiment Classes

Sentiment	Total
Positive	2688
Neutral	4666
Negative	2646

After determining the sentiment value of each opinion sentence, the opinion sentences are formed into a dataset using the Attribute-Relation File Format (ARFF)[29] as the input for classifying data with WEKA. The tweet data were then classified or tested for accuracy using the NBC machine learning algorithms and SVM with WEKA version 3.8.3 software.

This study uses the 10-fold cross-validation method for the process of classifying or testing tweet data. In this process, the data are divided into 10 parts with 9/10 parts used for the training process and 1/10 parts used for the testing process. Iteration takes place 10 times with variations in training and testing data using a combination of 10 parts of data. Table 5 displays a comparison of the results from the NBC machine learning algorithm with SVM.

Table 5: Comparison of Classification Results

Naïve Bayes Classifier (NBC)					
Tokenizer	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**TP Rate (%)**	**TN Rate (%)**
Alphabetic	49.94	52.5	49.9	51.4	50.5

https://doi.org/10.31436/iiumej.vxxixx.xxxx

Character N-gram	51.81	54.4	51.8	53.7	51,7
Unigram	50.11	52.6	50.1	53.3	48.4
Bigram	50.51	61.2	50.5	79.7	25.3
Trigram	55.98	68.5	56	21	20.3
N-gram	44.94	51.6	44.9	64.2	46
Word	50.11	52.6	50.1	53.3	48.4
Support Vector Machine (SVM)					
Tokenizer	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**TP Rate (%)**	**TN Rate (%)**
Alphabetic	79.02	79	79	73.8	70.3
Character N-gram	58.47	59.20	58.50	55.70	64.40
Unigram	78.9	78.9	78.9	73.2	70.8
Bigram	69.21	72.1	69.2	51.5	48
Trigram	61.21	71.7	61.2	30.1	31
N-gram	77.82	77.8	77.8	72.6	69.4
Word	78.9	78.9	78.9	73.2	70.8

*) The Precision and Recall values are the average values of positive class values and negative classes.

The information in Table 5 enables a comparison of the accuracy, precision, recall, TP rate, and TN rate values for each trial carried out with the NBC machine learning algorithm and SVM. The columns contain the tokenization data used in this study while the rows contain the accuracy, precision, recall, TP rate, and TN rate values for each trial conducted. The process from data preprocessing to the determination of the sentiment class produced the dataset of this research, which was then used as the input in the classification process. The classification process was carried out with WEKA Machine Learning using the NBC machine learning algorithm and SVM. The classification test process with 7 tokenizations produced values for accuracy, precision, recall, TP rate, and TN rate for each trial.

https://doi.org/10.31436/iiumej.vxxixx.xxxx

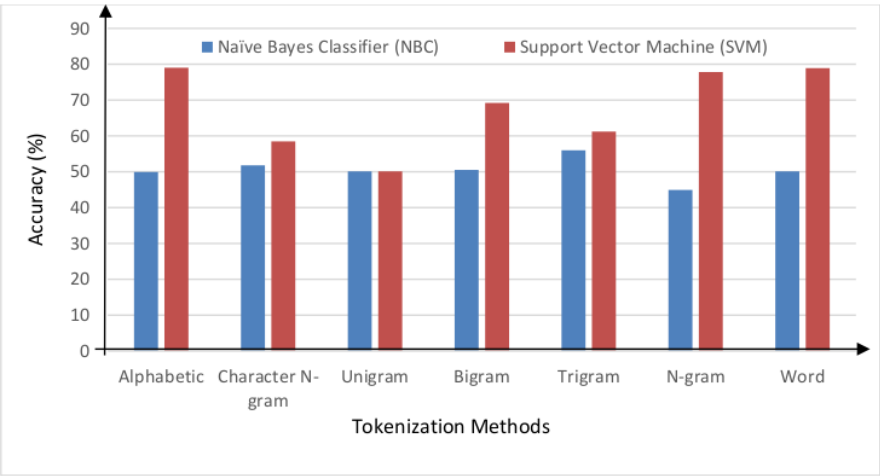

Fig. 5. Accuracy level of different machine learning alghoritm and tokenization methods. The blue and orange charts correspond to the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM), respectively.

From Figure 5, we can see that accuracy was tested in this study with two machine learning algorithms, namely NBC and SVM, and the 7 tokenizations of Alphabetic Tokenizer, Character N-gram Tokenizer, Unigram, Bigram, Trigram, N-gram, and Word Tokenizer. Accuracy was one of the main parameters in the assessment of the sentiment analysis model used in this study. The formula for the value of accuracy was the amount of data that were successfully classified according to the class of sentiment for the entire amount of data classified. Therefore, the greater the amount of data that were correctly classified according to the sentiment class, the higher the accuracy value. The highest accuracy value was obtained with respect to the combination of the SVM and Alphabetic Tokenization machine learning algorithms, which had an accuracy value of 79.02%. In this study, machine learning methods such as the SVM algorithm produced the highest accuracy because they work by recognizing word patterns. This machine learning algorithm is capable of easily recognizing and memorizing word patterns for a certain sentiment class in an opinion sentence. Yet while it is easy to classify sentiment data correctly using these methods, alphabetic tokenization can improve accuracy by breaking a sentence into words, which enables the easy classification of sentences with sentiments. The lowest accuracy value in this study was obtained for the NBC machine learning algorithm with N-gram tokenization, which yielded an accuracy value of 44.94%.

https://doi.org/10.31436/iiumej.vxxixx.xxxx

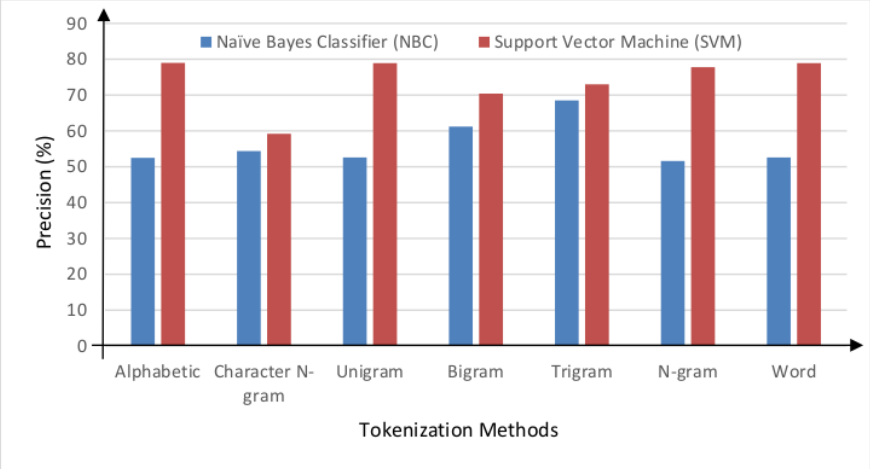

Fig. 6. Precision level of different machine learning alghoritm and tokenization methods. The blue and orange charts correspond to the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM), respectively.

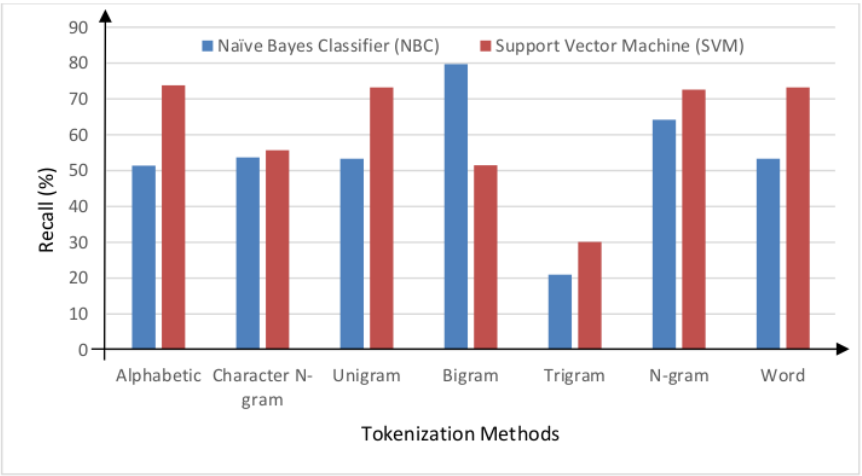

Fig. 7. Recall level of different machine learning alghoritm and tokenization methods. The blue and orange charts correspond to the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM), respectively.

Figure 6 shows that the highest Precision value of 79% was obtained by the SVM machine learning algorithm with alphabetic tokenization, while the lowest precision value of 55.1% came from the NBC machine learning algorithm with N-gram tokenization. In the Figure 7, we can see that highest Recall value of 79% was obtained with the SVM machine learning algorithm and alphabetic tokenization, while the lowest Recall value of 51.6% was obtained with the NBC machine learning algorithm with N-gram tokenization. The high precision values were obtained because the precision value formula was based on the number of positive classes

https://doi.org/10.31436/iiumej.vxxixx.xxxx

that were correctly classified as a positive class divided by the total data classified as a positive class, whereas the recall value formula consisted of the number of positive classes that were correctly classified as positive classes divided by the number of actual positive classes.

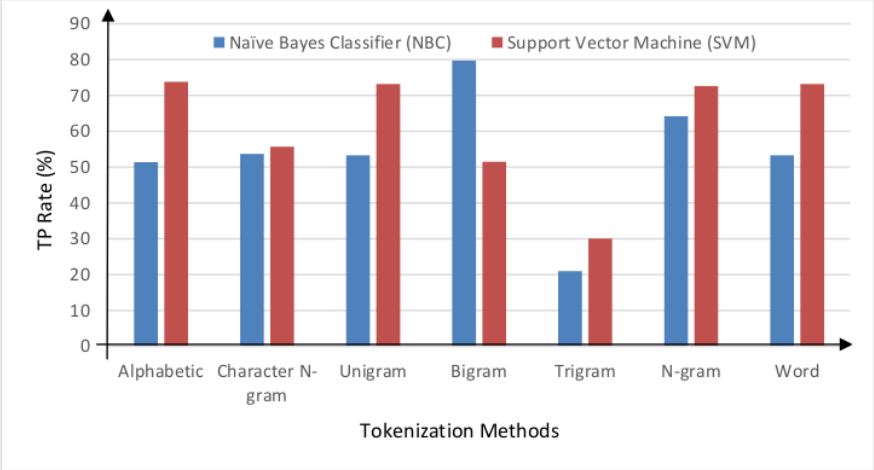

Fig. 8. TP Rate level of different machine learning alghoritm and tokenization methods. The blue and orange charts correspond to the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM), respectively.

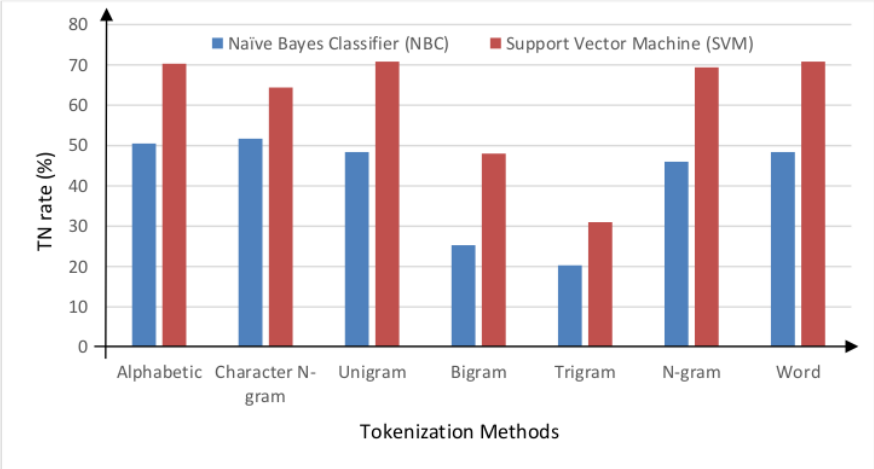

Fig. 9. TN Rate level of different machine learning alghoritm and tokenization methods. The blue and orange charts correspond to the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM), respectively.

Figure 8 displays the highest and lowest TP Rate obtained using the NBC machine learning algorithm. The TP Rate is a value denoting the amount of positive tweet data that were correctly classified according to the sentiment class, which in this case was positive. In contrast, the highest and lowest TN values were obtained using the SVM machine learning algorithm as

shown in Figure 9. The TN Rate indicates the value of negative tweet data that were correctly classified according to the sentiment class, which in this case was negative.

From the research carried out, it can be seen that the model constructed delivered the greatest accuracy when using a combination of the SVM machine learning algorithm and N-gram tokenization, while the lowest accuracy value was obtained when testing using a combination of the NBC machine learning algorithm with Trigram tokenization. The accuracy results produced were quite good; however, the model still made a number of mistakes when the classification process of the dataset with the distribution of sentiments was not as balanced as this study intended. The use of datasets with imbalanced distribution will lead to the incorrect classification of minority class data as majority class data[30], which results in a large value difference because most classifiers manage to correctly classify the majority class compared to the minor class[31].

5. CONCLUSIONS

From the series of studies conducted, we can conclude that the Sentiment Analysis model built was suitable for use in determining the sentiment of public opinion on Twitter with respect to the 2019 Indonesian presidential candidates. The study aimed to test and determine which machine learning algorithms were suitable for the classification of public opinion on Twitter, and also to test 7 suitable tokenizations and produce high accuracy when combined with the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) machine learning algorithms. The sentiment analysis revealed that there was much negative public sentiment on Twitter aimed at the 2019 Indonesian presidential candidates. The greatest accuracy value was obtained when using a combination of the SVM machine learning algorithm and alphabetic tokenization, which yielded an accuracy value of 79.02%. The lowest accuracy value in this study was obtained for the NBC machine learning algorithm with N-gram tokenization, which had an accuracy value of 44.94%. This study has therefore demonstrated that the SVM machine learning algorithm produces higher accuracy compared to the NBC machine learning algorithm. It is suggested that further research should endeavor to use more data and real-time data from both Twitter and other social media sites such as Facebook and YouTube.

Funding: This research was funded by Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876, and the Fundamental Research Grant Scheme (FRGS) Vot 5F073 supported under the Ministry of Education Malaysia. The work is partially supported by the SPEV project, University of Hradec Kralove, FIM, Czech Republic (ID: 2103-2019), “Smart Solutions in Ubiquitous Computing Environments” and by the project of excellence 2019/2205, Faculty of Informatics and Management, University of Hradec Kralove.

ACKNOWLEDGEMENT

This research was partially funded by Universitas Muhammadiyah Ponorogo through the Institute for Research and Public Services (LPPM) under the contract no 115/VI.4/PN/2018. We are also grateful for the support of Ph.D. students Sebastien Mambou and Michal Dobrovolny in consultations regarding application aspects.

REFERENCES

[1] KPU, "KPU - Portal Publikasi Pemilihan Umum 2019," 2018. [Online]. Available: https://infopemilu.kpu.go.id/. [Accessed: 27-Jun-2020].

[2] S. Aslam, "Instagram by the Numbers: Stats, Demographics & Fun Facts," *Omnicore Agency*, 2020. [Online]. Available: https://www.omnicoreagency.com/instagram-statistics/. [Accessed: 27-Jun-2020].

[3] L. Januru, "ANALISIS WACANA BLACK CAMPAIGN (KAMPANYE HITAM) PADA PILPRES TAHUN 2014 DI MEDIA KOMPAS, JAWA POS DAN KEDAULATAN RAKYAT," *NATAPRAJA*, vol. 4, no. 2, pp. 181–194, 2016.

[4] N. Indurkhya and F. J. Damerau, *Handbook of Natural Language Processing*, 2nd ed. Chapman & Hall/CRC, 2010.

[5] B. Wagh, S. J. V., and W. N. R., "Sentimental Analysis on Twitter Data using Naive Bayes," *Ijarcce*, vol. 5, no. 12, pp. 316–319, 2016.

[6] D. Yuret and F. Türe, *Learning morphological disambiguation rules for Turkish*. 2006.

[7] N. Zainuddin and A. Selamat, "Sentiment analysis using Support Vector Machine," in *I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings*, 2014, no. September, pp. 333–337.

[8] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, 2018.

[9] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Proc. 7th Int. Conf. Lang. Resour. Eval. Lr. 2010*, no. December, pp. 1320–1326, 2010.

[10] O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," 2015.

[11] Z. B. Nezhad and M. A. Deihimi, "A combined deep learning model for Persian Sentiment Analysis," *IIUM Eng. J.*, vol. 20, no. 1, pp. 129–139, 2019.

[12] N. Zainuddin, A. Selamat, and R. Ibrahim, *Improving Twitter Aspect-Based Sentiment Analysis Using Hybrid Approach*, vol. 9621. 2016.

[13] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," *Appl. Intell.*, vol. 48, Dec. 2017.

[14] N. S. Purohit, A. B. Angadi, M. Bhat, and K. C. Gull, "Crawling through web to extract the data from Social networking site-Twitter," *2015 Natl. Conf. Parallel Comput. Technol. PARCOMPTECH 2015*, 2015.

[15] "R: What is R?" [Online]. Available: https://www.r-project.org/about.html. [Accessed: 27-Jun-2020].

[16] X. Zhang *et al*., "Encyclopedia of Machine Learning," 2011, pp. 456–457.

[17] F. S. Utomo, N. Suryana, and M. S. Azmi, "Stemming impact analysis on Indonesian Quran translation and their exegesis classification for ontology instances," *IIUM Eng. J.*, vol. 21, no. 1, pp. 33–50, 2020.

[18] A. Lutfiatun, A. Novitasari, and A. Helfiyana, "Bahasa Alay Pada Chating Di Medsos Remaja Millenial (Bahasa Alay Vs Remaja Millenial)," *Pros. SENASBASA*, pp. 34–41, 2018.

[19] "Pencarian - KBBI Daring." [Online]. Available: https://kbbi.kemdikbud.go.id/. [Accessed: 14-Mar-2020].

[20] "Weka_tokenizers: R/Weka Tokenizers in RWeka: R/Weka Interface." [Online]. Available: https://rdrr.io/cran/RWeka/man/Weka_tokenizers.html. [Accessed: 14-Mar-2020].

[21] M. Jabreel and A. Moreno, "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets," *Appl. Sci.*, vol. 9, p. 1123, Mar. 2019.

[22] C. Musto, G. Semeraro, and M. Polignano, "A comparison of lexicon-based approaches for sentiment analysis of microblog," *CEUR Workshop Proc.*, vol. 1314, pp. 59–68, 2014.

[23] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *Integer J. Maret*, vol. 1, no. 1, pp. 32–41, 2017.

[24] "Machine Learning Group - Department of Computer Science: University of Waikato." [Online]. Available: https://www.cs.waikato.ac.nz/research/research-groups/machine-learning-group. [Accessed: 27-Jun-2020].

[25] M. Wiley and J. F. Wiley, *Advanced R statistical programming and data models: Analysis, machine learning, and visualization*. Apress Media LLC, 2019.

[26] "Confusion Matrix for Your Multi-Class Machine Learning Model." [Online]. Available: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826. [Accessed: 27-Jun-2020].

[27] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina, V. Martinez-Hernandez, V. Sanchez, and H. Perez-Meana, "A Web Scraping Methodology for Bypassing Twitter API Restrictions," pp. 1–7, 2018.

[28] K. Haniewicz, W. Rutkowski, M. Adamczyk, and M. Kaczmarek, *Towards the Lexicon-Based Sentiment Analysis of Polish Texts: Polarity Lexicon*. 2013.

[29] "Attribute-Relation File Format (ARFF)." [Online]. Available: https://www.cs.waikato.ac.nz/~ml/weka/arff.html. [Accessed: 27-Jun-2020].

[30] "Handling imbalanced datasets in machine learning - Towards Data Science." [Online]. Available: https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28. [Accessed: 27-Jun-2020].

[31] R. Longadge and S. Dongre, "Class Imbalance Problem in Data Mining Review," *Int. J. Comput. Sci. Netw.*, vol. 2, May 2013.

THE IMPLEMENTATION OF THE MACHINE LEARNING ALGORITHM FOR THE SENTIMENT ANALYSIS OF INDONESIA'S 2019 PRESIDENTIAL ELECTION

ORIGINALITY REPORT

24%	15%	13%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

6%

★ www.ijcaonline.org

Internet Source

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		