Classification of Javanese Language Level on Articles Using Multinomial Naive Bayes and N-Gram Methods

To cite this article: A P Ardhana et al 2019 J. Phys.: Conf. Ser. 1306 012049

View the article online for updates and enhancements.
Classification of Javanese Language Level on Articles Using Multinominal Naïve Bayes and N-Gram Methods

A P Ardhana, D E Cahyani, and Winarno
Informatics Department of Mathematics and Natural Sciences Faculty of Sebelas Maret University, Indonesia
E-mail: adhikapri_ardhana@student.uns.ac.id, denis.eka@staff.uns.ac.id, win@staff.uns.ac.id

ABSTRACT. Javanese language articles can now be found in electronic media. But for people who are just learning the Javanese language, they have difficulty in knowing the language level that is contained in the article because there is no explanation regarding the type of language level. For this reason, it is necessary to classify the language level in Javanese language based on the article, where the Javanese language is divided into 4 levels, namely ngoko, ngoko alus, krama madya, and krama alus. Before beginning the classification process, the data must be preprocessed. One of the steps in preprocessing method is stemming, which is used to convert the affixes in words into basic words. The Javanese stemming process in this research refers to the Indonesian stemming rules based on the adjusted Nazief-Adriani algorithm. The extraction feature process was carried out using N-Gram with n = 2,3,4 (bigram, trigram, and quadgram). After finishing the preprocessing method, the classification process then was executed using Multinomial Naïve Bayes method. In the classification process, there often exist problems related to imbalance data between categories. To overcome this problem, SMOTE resampling method is utilized to balance the data. The classification with N-Gram variations accompanied by SMOTE and using stemming results to maximum accuracy of 67.01% at n = 2, then for the highest precision was at N-Gram n = 4 with condition not using stemming but using SMOTE at value 72.67%. The highest recall value was obtained in two conditions, namely at N-Gram n = 2 either using or not using SMOTE and use stemming, which is 67.00%. From the results of this study, it can be concluded that stemming rules which are adapted from Nazief-Adriani algorithm with the addition of stemming steps and affix list from Javanese language expert can be implemented properly for Javanese language stemming process. And the classification of the Javanese Language level using Multinomial Naïve Bayes and N-Gram methods can result in good enough accuracy, precision, and recall.

1. Introduction
Javanese language is a language which are spoken by many natives in Indonesia. According to Alva, based on the number of speakers, Javanese is ranked 11th in the world and is spoken in 5 countries other than Indonesia [1]. But, Javanese language is now in an alarming condition. Subroto [2] said that the younger generation does not like learning and using Javanese language because it has some spoken language levels according to their speaking opponents, namely ngoko, ngoko alus, krama madya and krama inggil [3]. Now, many people are trying to make Javanese language more popular to the younger generation, one the other way is write it in electronic media such as www.penjebarsangat.com and www.bantulkab.go.id/pawartos. But unfortunately, the article written
on the website does not include the Javanese language level used. So, we need a system that can classify Javanese language levels.

Before beginning the classification process, the data must be preprocessed. One of the steps in preprocessing method is stemming, which is used to convert the affixes in words into basic words. The Javanese stemming process in this research refers to the Indonesian stemming rules based on the adjusted Nazief-Adriani algorithm [4]. Previous research discussing the classification of Javanese language documents did not focus on making Javanese stemming rules[5]. In that research, the making of Javanese stemming rules did not refer to a certain rule.

The problem that often occurs in the data collection process is the limited data in certain categories, which results to resulting to unbalanced data. For this reason, a technique is needed to balance the data. There are various ways used to balance data, one of which is the Synthetic Minority Over-sampling Technique (SMOTE) method.

This paper discusses the classification language levels in Javanese language articles using the Multinomial Naïve Bayes method by adding the n-gram feature. Because of the data imbalance, the SMOTE process will be carried out. This paper also developed rules for the process of stemming the Javanese language by referring to the rules of Nazief-Adriani.

This paper is structured as follows: section 2 explains about Text Mining and Text Preprocessing. Section 3 explains about Javanese Stemming. Section 4 explains about N-Gram, Section 5 explains about Synthetic Minority Over-Sampling Technique (SMOTE). Section 6 explains about Multinomial Naïve Bayes. Section 7 explains about Data Processing. Section 8 explains about Result and Evaluation and section 9 concludes this paper.

2. Text Mining and Text Preprocessing
Text mining generally refers to the process of extracting information from unstructured text documents. Text mining can be defined as the discovery of new and previously unknown information by a computer, which automatically extracts information from different unstructured text sources. The key to this process is to combine information which are successfully extracted from various sources [6].

But before processing the text mining, preprocessing needs to be done. Preprocessing is a stage to prepare the text into data that will undergo further processing. The stages of text preprocessing: case folding is used to change all letters in an article into lowercase and delete the other characters except letters. Normalization is used to change the acronym to its extension and change the non-standard word to standard word. Tokenization is used to convert string into token or words. Stop word Removal is used to delete meaningless word. And the last is Stemming, which is used to convert the affixes in words into root words [7].

3. Javanese Stemming
Stemming is the stage of looking for the root word of each given word. The Javanese stemming process in this research refers to the Indonesian stemming rules based on the adjusted Nazief-Adriani algorithm, because there is no Javanese stemmer rule. In a previous research discussing Javanese language, classification did not focus on making Javanese stemming. In that research, affix lists (infix, suffix and prefix) have been set up in Javanese. But the stemming rules do not refer to certain rules and the affix list in the research is subjective.

4. N-Gram
N-gram is a method to cut a sentence into words. The n-gram extraction feature model is formed based on the frequency of the n-gram words that appear in the document [8]. The steps of N-Gram method are:
1. Dividing the text into tokens that only consist of letters.
2. Determining the value of n for each group of words with n = 2, 3, and 4.
3. Cutting each sentence into words as many as n.
4. Entering each piece of character into the table. Each piece of word has an initial value of 1. When there are pieces of the same word, the value in the table increases by 1.

5. **Synthetic Minority Over-Sampling Technique (SMOTE)**

Unbalanced data between categories can affect the result’s accuracy, so resampling must be done so that the data can be balanced. One of the methods in resampling is Synthetic Minority Over-sampling Technique (SMOTE). The SMOTE technique is an over-sampling approach in which the minority class is over-sampled by creating “synthetic” examples rather than by over-sampling with replacement. Minority classes will later be oversampled by taking each sample of minority classes and introducing artificial examples along line segments that connect each / all of the neighboring minority classes. Depending on the amount of oversampling required, neighbors from the nearest neighbor k are randomly selected [9].

The way SMOTE works is: in each instance $x_i$ in the minority class, SMOTE looks for its closest neighbor and one neighbor is randomly chosen as $x'$. Then a random value between $[0,1]$ $\delta$ is generated. New artificial samples $(x_{\text{new}})$ are made based on:

$$x_{\text{new}} = x_i + (x' - x_i) \delta$$  \hspace{1cm} (1)

6. **Multinomial Naïve Bayes**

Naïve Bayes is a probability statistical method based on the application of the Bayes Theorem with a strong independent assumption to predict the class of a document based on its probability [10]. Bayes’ Theorem is expressed by the formula:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$ \hspace{1cm} (2)

Where $X$ is data with unknown class, $C$ is the hypothesis data in a specific class, $P(C|X)$ is probability hypothesis based on condition (posterior probability), $P(C)$ is Probability hypothesis (prior probability), $P(X|C)$ is Probability based on condition in hypothesis and, $P(X)$ is Probability of hypothesis $X$.

Multinomial Naïve Bayes is one of the specific methods of Naive Bayes [10]. The probability of a document $d$ for class $C$ can be calculated with the following formula:

$$P(C|\text{term dokumen } d) = P(X_1|C) \times P(X_2|C) \times \ldots \times P(X_n|C) \times P(C)$$ \hspace{1cm} (3)

Prior probability class $C$ is determined by the formula:

$$P(C) = \frac{N_C}{N}$$ \hspace{1cm} (4)

The probability of the $n$ word is determined by:

$$P(X_n|C) = \frac{N_{x_n,C} + \alpha}{N(C) + V} \hspace{1cm} n = \{1, \ldots, n\}$$ \hspace{1cm} (5)

While the MNB formula used with the N-Gram feature extraction is as follows:

$$P(X_n|C) = \frac{\Sigma tf(x_n, d \in C) + \alpha}{\Sigma N_{d \in C} + V}$$ \hspace{1cm} (6)

7. **Data Processing**

7.1 **Data on Javanese Language Level**

The data in this research were taken from www.penjebarsemangat.com, www.bantulkab.go.id/pawartos, http://jogjatv.tv/category/berita/pawartos-ngayogyakarta/, http://www.pawartosjowo.com/, http://www.djakalodang.co.id/. From that source, we can get 200 articles which are categorized into 4 levels of language, namely “ngoko”, “ngoko alus”, “krama madya” and “krama inggil”. Distribution of the data is presented in this table:
Table 1. Data distribution

| Category         | Ngoko | Ngoko Alus | Krama Madya | Krama Alus |
|------------------|-------|------------|-------------|------------|
| Total            | 76    | 16         | 25          | 83         |

7.2 Text Preprocessing

In text preprocessing, there are 5 stages, namely: case folding, normalization, tokenization, stop word removal and stemming. Case folding is used to change all letters in an article into lowercase and delete the other characters except letters. Normalization is used to change the acronym to its extension and change the non-standard word to standard word. Tokenization is used to convert string into token or words. Stop word removal is used to delete meaningless word. Stemming is used to convert the affixes in words to root words.

The process of stemming and stopword removal used the modified version of Nazief-Adriani rules through advice from a Javanese language expert, and based on a book called Paramasastra Jawa (in which the affix list was implemented) [11]. Examples of the application of the stemming process are, for example, there is a sentence "rampung acara kemah nang bumi perkemahan wanagama". Then each word will be processed according to the stemming rules discussed earlier:

1. Select the word that will be stemmed.
2. Checking the word in the dictionary whether the word includes the basic word or not. If it is in the dictionary, then it is finished. If not, then the process will enter the prefix checking phase.
   - rampung → rampung (the process is stopped, because word *rampung* exists in the dictionary).
   - perkemahan → perkemahan (go to prefix step because the word did not exist in the dictionary).
3. If the word has a prefix, cut the word according to the prefix rules that already exist and then check if the word already exists in the dictionary. If the word is in the dictionary, the process stops. If not, then proceed to the infix rule.
   - perkemahan → kemahan (go to suffix step because the word did not exist in the dictionary).
4. If the word contains infix, cut it according to the infix rule and then check if the word already exists in the dictionary. If the word is in the dictionary, then return the word. If not, then proceed to the suffix rule.
   - kemahan → kemah (the process is stopped, because word *kemah* available in dictionary).
5. If the word contains a suffix, cut the existing suffix rule and then check in the dictionary. If the word already exists in the dictionary, then stop and return it. If not, then change the word to how it was before it was processed in the stemming rule.
   - bumi → bumi (finished, because the word has no infix and the word bumi is in the dictionary).

The full text preprocessing process is shown in the following figure:
Figure 1. Example of text preprocessing

7.3 Feature Extraction with N-Gram

Feature Extraction process with N-Gram aims to calculate the weight of each word in the article. The results of the weighting will be used in the classification process. Examples of the weighting process with N-Gram while n=2 are:

Table 2. Example of feature extraction with n=2

| Data                     | N-Gram’s Weight |
|--------------------------|-----------------|
| rampung acara kemah bumi| 1               |
| kemah bumi               | 1               |
| bumi kemah               | 1               |
| kemah wanagama           | 1               |
| wanagama rudi            | 1               |
| rudi langsung            | 1               |
| langsung mulih           | 1               |
| mulih menyang            | 1               |
| menyang omah             | 1               |
| omah simbah              | 1               |
7.4 Resampling
Because the data is not balanced, it is necessary to balance the data with oversampling method, namely Synthetic Minority Over-sampling Technique (SMOTE). Results from SMOTE can be seen in the table below:

| Category       | Total original data | Data after SMOTE |
|----------------|---------------------|------------------|
| Ngoko          | 16                  | 83               |
| Ngoko Alus     | 76                  | 83               |
| Krama Madya    | 25                  | 83               |
| Krama Inggil   | 83                  | 83               |

7.5 Classification
After the data is balanced with SMOTE and calculated with N-Gram (bigram, trigram, and quadgram), the next step is to classify with Multinomial Naïve Bayes (MNB) method. MNB classification uses two data, namely training data and testing data. Training data is used as learning material in the testing process to determine a data entered in a certain category.

Calculation stage of the classification process are:
1. Calculating the prior value of each category using equation ().
2. Calculating the probability of the nth word of data in the article using equation () or ().
3. Permitting the probability of an article entering a category with equation ()
4. Determining the document class by selecting the highest probability value.

The classification process in this study consists of two conditions, namely stemming and without stemming. In each stage, data were also carried out in 2 treatments: data through the SMOTE stage and data not through the SMOTE stage.

8. Result and Evaluation
The objective of this section is to compare the maximum results between all conditions. This comparison is taken from each of the maximum precision, recall, and accuracy values for each condition. Then in the dividing data for training data and testing data for the classification, using Cross Validation. Feature extraction process used several variations of the N-Gram type, namely n = 2,3,4, while the data sharing process used k-fold parameter = 2,3,4,5,6,7,8,9,10, thus, resulting to:

First is the result of precision. From table 4 shows that with condition data not going through the stemming stage and the amount of data balanced with SMOTE, the highest precision value is at n-gram = 4, which is 72.67%.
Table 4. Precision Comparison

| N-Gram                                      | Precision |
|---------------------------------------------|-----------|
| 2 (Stemming and SMOTE)                      | 65.86%    |
| 3 (Stemming and SMOTE)                      | 64.67%    |
| 4 (Stemming and SMOTE)                      | 68.67%    |
| 2 (Stemming and without SMOTE)              | 67.44%    |
| 3 (Stemming and without SMOTE)              | 60.00%    |
| 4 (Stemming and without SMOTE)              | 58.50%    |
| 2 (Without Stemming and SMOTE)              | 68.00%    |
| 3 (Without Stemming and SMOTE)              | 60.00%    |
| 4 (Without Stemming and SMOTE)              | 58.50%    |
| 2 (Without Stemming and without SMOTE)      | 68.00%    |
| 3 (Without Stemming and without SMOTE)      | 61.50%    |
| 4 (Without Stemming and without SMOTE)      | 57.50%    |

Table 5. Recall Comparison

| N-Gram                                      | Recall    |
|---------------------------------------------|-----------|
| 2 (Stemming and SMOTE)                      | 67.00%    |
| 3 (Stemming and SMOTE)                      | 46.33%    |
| 4 (Stemming and SMOTE)                      | 38.57%    |
| 2 (Stemming and without SMOTE)              | 67.00%    |
| 3 (Stemming and without SMOTE)              | 52.33%    |
| 4 (Stemming and without SMOTE)              | 46.33%    |
| 2 (Without Stemming and SMOTE)              | 64.22%    |
| 3 (Without Stemming and SMOTE)              | 47.00%    |
| 4 (Without Stemming and SMOTE)              | 40.33%    |
| 2 (Without Stemming and without SMOTE)      | 64.22%    |
| 3 (Without Stemming and without SMOTE)      | 51.50%    |
| 4 (Without Stemming and without SMOTE)      | 46.33%    |

The results from table 5 shows that with condition either using or not using SMOTE and using stemming, the highest recall value is at n-gram = 2, which is 67.00%.

Then, the results from table 6 show that with condition data going through the stemming stage and the amount of data balanced with SMOTE, the highest accuracy value at n-gram = 2, which is 67.01%.

Table 6. Accuracy Comparison

| N-Gram                                      | Accuracy  |
|---------------------------------------------|-----------|
| 2 (Stemming and SMOTE)                      | 67.01%    |
| 3 (Stemming and SMOTE)                      | 46.49%    |
| 4 (Stemming and SMOTE)                      | 38.60%    |
9. Conclusion

From the results of this research, it can be concluded that stemming rules referring to Nazief-Adriani rules with the addition of stemming rule and affix list from Javanese linguists can be implemented well for the Javanese language stemming process. The resampling process using SMOTE gives an increase for the classification results. With condition data not going through the stemming stage and using SMOTE, the highest precision value is at N-Gram = 4, which is 72.67%. The highest recall value with the percentage of 67.00% was found in two conditions, namely at N-Gram = 2 with condition either using or not using SMOTE and using stemming. Then, with condition data going through the stemming stage and the amount of data balanced with SMOTE, the highest accuracy value is at N-Gram = 2, which is 67.01%.

References

[1] Alva, A., *Makin Mendunia, Berbanggalah Jadi Orang Jawa! Ini 7 Fakta Unik Bahasa Jawa*. 2018.
[2] Subroto, D.E., M. Rahardjo, and B. Setiawan, *Endangered krama and krama Inggil varieties of the Javanese language*. Linguistik Indonesia, 2008. 26(1): p. 89-96.
[3] Sasangka, S.S.T.W. and Y. Maryani, *Unggah-ungguh bahasa Jawa*. 2004: Yayasan Paramalingua.
[4] Nazief, B. and M. Adriani, *Confix Stripping: Approach to Stemming Algorithm for Bahasa Indonesia*. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta, 1996.
[5] Yosnaningsih, Y.V., *Klasifikasi Dokumen Bahasa Jawa Menggunakan Metode Naïve Bayes*. Sanata Dharma University, 2015.
[6] Tan, A.-H. *Text mining: The state of the art and the challenges*. in *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*. 1999. sn.
[7] Efendi, R. and J. MilaSari, *Klasifikasi dokumen berbahasa indonesia Menggunakan naive bayes classifier*. Journal Research Computer Science & Application, 2012. 1(1).
[8] Cavnar, W.B. and J.M. Trenkle, *N-gram-based text categorization*. Ann arbor mi, 1994. 48113(2): p. 161-175.
[9] Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique*. Journal of artificial intelligence research, 2002. 16: p. 321-357.
[10] Simanullang, J.W., A. Adiwijaya, and S. Al Faraby, *Klasifikasi Sentimen Pada Movie Review Dengan Metode Multinomial Naïve Bayes*. eProceedings of Engineering, 2017. 4(2).
[11] Padmosoekotjo, S., *Paramasastra Jawa*. 1986: Citra Jaya Murti.