Effectiveness of the News Text Classification Test Using the Naïve Bayes’ Classification Text Mining Method

Yi Ying1*, T N Mursitama2, Shidarta3 and Lohansen4

1,4Chinese Department, Faculty of Humanities, Bina Nusantara University, Jl. Kemanggisan Ilir III No. 45, Indonesia
2International Relations Department, Faculty of Humanities, Bina Nusantara University, Jl. Kemanggisan Ilir III No. 45, Indonesia
3Business Law Department, Faculty of Humanities, Bina Nusantara University, Jl. Kemanggisan Ilir III No. 45, Indonesia

*yi_ying@binus.edu

Abstract. This study used a quantitative approach created by the author to describe the accuracy of the classification machine learning for news text. In this report we compare the results of the accuracy values obtained using the Naïve Bayes method with other methods to see the effectiveness of the method used. The resulting accuracy value has not reached its maximum and thus it could still be restructured and re-evaluated into a better model. In this case the writer tried to increase the precision value produced in order to make this machine able to predict news that contains sarcasm through modification of the set threshold value. After the threshold value was changed to 0.3, the accuracy value decreased to 61% but the precision value increased to 77% and the error value in the prediction of false positive headlines also increased significantly and only produced 89 errors in predicting sarcasm. The ROC Curve test showed that this machine learning model could still be improved by trying other text preprocessing methods such as the bigrams, tidyttext, lemmatization methods, so that the machine will become smarter at predicting the resulting vectors and increase the value of precision and accuracy obtained.

1. Introduction

News has become an institution used to distribute the latest information needed by the people. Through various agencies, news is conveyed using media such as Online Media, Television, Newspapers, Radio, and various other media. In general, the news delivered in the media consists of several categories such as political, sports, economic, health, and so forth. However, at present classifying the news into these categories is still done manually, meaning that when uploading a news item, the uploader must first know the whole content of the news to be uploaded which is then put in the right category. This is very inconvenient for news uploaders who are dealing with large numbers of news items. Therefore, it is necessary to have a web-based system which can classify the news automatically according to existing news categories so that it can help the news uploaders when uploading such material. To overcome these problems, a machine learning method that can clarify news content types has been developed. This clarification method can be carried out using a variety of...
clustering models in the machine learning model, and in preprocessing data on news texts, the method used is text mining. Text mining refers to the general process of extracting patterns or knowledge contained in unstructured text documents. This can be seen as an extension of data mining or knowledge discovery from a structured (letter and nominal) database. The text has become one of the most natural forms of data stored and therefore text mining is believed to have higher commercial potential than data mining. In fact, recent research shows that 80% of company information is presented in text documents. However, text mining also comprises a large number of more complex tasks (compared to data mining) because it involves dealing with text data that are inherently unstructured and vague. Text mining is a multidisciplinary field, which involves information retrieval, text analysis, information extraction, and grouping. The text mining classification method is one of the techniques that can be used to make classifications [1]. Text mining is a variation of data mining that tries to identify interesting patterns in a large collection of textual data. Besides classification, text mining is also used to deal with issues related to clustering, information extraction, and information retrieval. Application text mining already used in psychiatry, which exploratory by nature, processing free speech or texts obtained from patients or physicians [2].

This study aimed to test the effectiveness of the NBCi (Naïve Bayes Classification) method for mandarin news texts, to calculate the value of accuracy, precision, recall obtained through the NBC method and to improve the values obtained. The goal was to ensure that Mandarin learners were also helped when working in media-related fields both in electronic and print media. Previous research has shown that Mandarin learners have difficulty in reading Mandarin news, especially economic news. This is because economic terms are perceived to be foreign. With the clarification of the news, it is hoped that learners will focus more on the type of news to be studied and better master the vocabulary associated with the news they read [2]. Machine learning is a form of AI (Artificial Intelligence) concept or often referred to as artificial intelligence that allows the system to learn from data rather than through explicit programming. However, machine learning is not a simple process. Machine learning uses various algorithms that repeatedly learn from data to improve, describe data, and predict results. As the algorithm digests training data, it is possible to produce a more appropriate model based on that data. [3].

2. Research Methods
The research method that will be used by the researchers is a quantitative research method, where the data obtained will be tested and further research will be carried out to produce a value that can be interpreted as information. This value will show how accurate the prediction of the NBC classification model that will be made. In conducting this quantitative research, the researchers will make a hypothesis first, then test the method to prove the hypothesis that is made beforehand and see the truth behind the hypothesis.

The hypothesis that the researchers will use is about the accuracy generated by the NBC method of Mandarin news text documents. This study uses data that is available and ready to be used in an NBC testing method. The data is taken from the Kaggle.com dataset, News Headlines Dataset for Sarcasm Detection, which is the result of scrapping data from the news portal website The Onion and Huffpost.

3. Result and Discussion

3.1. Literature Review
There are two stages in the NBC (Naïve Bayes Classification) method text classification process, namely the training phase and the classification stage. At the training stage, the process of analyzing the sample documents is in the form of vocabulary selection, i.e. words that may appear in the collection of sample documents that can as far as possible become representative documents. Next is
determining the prior probabilities for each category based on the sample documents. At the classification stage the category value of a document is determined based on the terms that appear in the documents classified.

According to tested 12 Indonesian news articles obtained from Kompas.com consisting of various categories. The categories were national, international, sports, science, education, economics, technology, entertainment, automotive, health, property, and travel news. 100 articles were taken for each category and thus the total number of articles used as data was 1200. The results of data mining using the NBC (Naïve Bayes Classification) and SVM (Support Vector Machine) methods, achieved an accuracy level of 82.2% and 88.1% respectively. It can be concluded that the NBC and SVM methods are able to classify Indonesian language news quite well in terms of the accuracy obtained [5].

The result of a study that compare the NBC method with the K-NN (K-Nearest Neighbors) method and the SVM method. The results of the study showed that the NBC, SVM and K-NN methods obtained an accuracy level of 90%; 92.5% and 50% respectively showed that SVM achieved good performance in the classification of Indonesian texts similar to what was achieved in the classification of English texts. SVM had a good ability to generalize in high dimensional feature spaces so that it did not require feature selection. Apart from that, SVM was strong and outperformed other conventional methods in all experiments. Surprising results came from the Naïve Bayes which showed the best performance among other conventional methods. Language structure only played a small role in this experiment because there was no difference in performance when SVM was applied to Indonesian text documents [6].

According finding from a research on text documents and academic documents using the NBC method to classify the texts showed a fairly high performance in the classification of text documents, both news and academic documents. The classification of news documents obtained higher accuracy with an accuracy value of 91% compared to 82% in academic documents. The use of unique words in the collection of training documents without filters could not achieve optimal performance. A word filter was tried using document frequency. It was found that a minimum filter of words appeared in four or five documents producing the highest accuracy results compared to other filters. However, it is not possible to determine the minimum value limit to be used as a reference when there are more documents. Researcher suggest that Hamzah suggested looking for a better technique for selecting word features to be used as a basis for classification, because it was found that a large number of words using all the unique words in the collection of documents did not provide the best classification results [7].

3.2 Naïve Bayes Classification

Naïve Bayes Classification (NBC) is a collection of classification algorithms based on the Bayes Theorem. NBC is not a single algorithm but a collection of several algorithms where they all have the same principle, that is, each pair of features classified does not depend on each other. Stage of making the Naïve Bayes Classification model:

1) Cross-Validating

In this stage the dataset will be divided into two parts first, namely the Data Train and Data Test.

a. Data Train

The Data Train contains 80% of the data that has been randomized and separated to be used as learning material from the Naïve Bayes Classification model that will be created. This data has been randomized using the set.seed function available in this R software.

b. Test Data
The test data will be used after the Naïve Bayes Classification model has been trained and has an algorithm formed through training using the previous data train model. This test data at the end of the modeling will be used again to compare the results obtained with the results of the model obtained, so that it can determine the accuracy of the model that has been created.

2) Bernoulli Convolution
The next step is to create the Bernoulli Convolution function to create a data train label and test data label to be detected by the machine. This label will be a sign of whether a news has a meaning of positive sarcasm or negative sarcasm which is changed into symbols 0 and 1.

3) Naïve Bayes Classification Modeling
After the data has been done cross validate and Bernoulli convolution. Then the data train is ready to be used for modeling using the Naïve Bayes Classification method. Through this modeling process, then we can find out what texts are in the data that have a meaning of positive sarcasm, and negative sarcasm through the classification of vectors that have been made previously through text preprocessing.

The author makes this NBC model named as sarcasm model. This model will be able to determine news that has the meaning of sarcasm both positive and negative. Then this model is entered into the test data that has previously been cross-validated to see the comparison of accuracy with the original test data. This test is performed using the "predict" function.

3.3 Confusion Matrix
After the model had been constructed and the testing completed, the last stage was evaluating the results of the model. The confusion matrix method was used for the evaluation forming the basis for observing the results of machine learning models and also the level of accuracy produced.

| Table 1. Confusion Matrix and Statistics |
|-----------------------------------------|
| Accurancy                               | 0.6671 |
| 95% CI                                  | (0.6418, 0.6918) |
| No Information Rate                     | 0.5386 |
| P-Value [Acc>NIR]                       | <2e-16 |
| Kappa                                   | 0.3291 |
| Mcnemar’s Test P-Value                  | 0.4871 |
| Sensitivity                             | 0.7016 |
| Specificity                             | 0.6269 |
| Pos Pred Value                          | 0.6870 |
| Neg Pred Value                          | 0.6429 |
| Prevalence                              | 0.5386 |
| Detection Rate                          | 0.3779 |
| Detection Prevalence                    | 0.5500 |
| Balanced Accuracy                       | 0.6643 |

The results of the Confusion Matrix show that the Sarcasm model above created using the Naïve Bayes method has succeeded in achieving an accuracy level of 66%, 70% withdrawal and 68% precision. There are 33% errors that occurred in predictions in this machine learning model. The results showed that out of a total of 630 news items in the test data, the machine predicted that 225 of
them were negative sarcasm, and 405 contained positive sarcasm. In fact there were 529 negative-sarcasm news items and the rest were positive.

In addition to the above, the Naïve Bayes model can be further developed to become smarter, in the sense that it will be able to predict news headlines that are sarcasm but were not predicted to be sarcasm. That's why in this case it could change the threshold entered into the model to increase the Precision value.

| Table 2. NBC Model threshold 0.3 |
|---------------------------------|
| Accurancy                       | 0.6193 |
| 95% CI                          | (0.5933, 0.6448) |
| No Information Rate             | 0.5386 |
| P-Value [Acc>NIR]               | 6.429e-10 |
| Kappa                           | 0.263 |
| McNemar’s Test P-Value          | <2.2e-16 |
| Sensitivity                     | 0.4111 |
| Specificity                     | 0.8622 |
| Pos Pred Value                  | 0.7769 |
| Neg Pred Value                  | 0.5564 |
| Prevalence                      | 0.5386 |
| Detection Rate                  | 0.2214 |
| Detection Prevalence            | 0.2850 |
| Balanced Accuracy               | 0.6367 |

The threshold in the first model was 0.5 (standard) which was then reduced to 0.3. Although the accuracy value decreased from the original 66% to 61%, there was a significant increase in the precision level which was initially 68% to 77% (an increase of almost 10%) which makes this model smarter in predicting news that contains sarcasm. The error value in the prediction of false positive headlines also increased significantly which only reduced 89 errors.

3.4 ROC Curve

The next step in testing the model used the ROC Curve. ROC is one of the additional evaluations used to assess whether this model has a good significance level or not. In the ROC test this time, it appeared that the results obtained were far below what they should have been. The ROC plot above shows that increasing the True Positive Rate or the level of precision produced, required a considerable effort, because there was a large gap in every increase from the Level True Positive to the Level of False Positive. Thus, this model is not considered to be suitable for use. In addition this model has a level of accuracy that was not so good, i.e. only around 60% -65% (although the threshold changed). This was possible because this model used preprocessing text mining with TM base library so the results obtained were not significant. A solution to that would be to try text preprocessing using other methods such as bigrams, tidytext or lematization methods to obtain better results.

4. Conclusions

Through the text mining process using the TM preprocessing method and Naïve Bayes Classification which was carried out on a news dataset from the news portal website The Onion and HuffPost, a machine was produced that is able to predict the sarcasm contained in a news item. This machine has an accuracy level of 66% and a precision value of 68%. It can be said that the resulting accuracy value has not reached the maximum level so that it could still be restructured and evaluated to produce a better model. In this case, the researcher has tried to increase the precision value produced in order to make this machine able to predict news that contains sarcasm through modification of the set threshold
value. When the threshold value was changed to 0.3, it resulted in an accuracy value that decreased to 61% but increased the precision value to 77%. The error value in the prediction of false positive headlines also increased significantly; the with prediction sarcasm errors only increasing to 89.

The ROC Curve test showed that this machine learning model could still be further improved by trying other text preprocessing methods such as bigrams, tidytext, lemmatization methods. This will make it possible to create a machine that is smarter at predicting the resulting vector and increasing the precision and accuracy values obtained.

References
[1] C C Aggarwal, C X Zhai 2012 Mining Text Data London: Springer
[2] A Abbe., et all 2016 Text mining applications in psychiatry: a systematic literature review International Journal of Methods in Psychiatric Research 25 (2) 86–100
[3] Y Ying, T N Mursitama 2019 Chinese Language Capability between Industry Needs and Shortage of Qualified Human Capital 2019. The 5th Friendly City International Conference (FCIC-5), Medan, 12 September 2019.
[4] Hurwitz, Judith dan Daniel Kirsch. 2018. Machine Learning for Dummies. IBM Limited Edition. (Online di : https://www.ibm.com/downloads/cas/GB8ZMQZ3)
[5] D Ariadi &. K Fithriasari 2015 Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer [Indonesian News Classification Using the Naive Bayesian Classification Method and Support Vector Machine with Confix Stripping Stemmer. Journal Sains & Seni ITS 4 (2) 2237-3520
[6] F. Wulandini & A. N. Nugroho 2009 Text Classification Using Support Vector Machine for Webmining Based Spation Temporal Analysis of the Spread of Tropical Diseases. International Conference on Rural Information and Communication Technology 2009. Retrieved from http://aswugroho.net/papers/ric2009_textclassification.pdf.
[7] A. Hamzah 2012 Klasifikasi Berita menggunakan Metode Naïve Bayes Classifier [News Classification uses the Naïve Bayes Classifier Method]. Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, Yogyakarta 3 November 2012. Retrieved from http://repository.akprind.ac.id/sites/files/conference-proceedings/2012/hamzah_15430.pdf

Acknowledgement
The authors gratefully acknowledge that the present research is supported by Ministry of Education Taiwan. The support is under the research grant Taiwan Studies Project of Year 2019.