Extraction Sentiment Analysis Using naive Bayes Algorithm and Reducing Noise Word applied in Indonesian Language

Aris Tri Jaka Harjanta¹, Bambang Agus Herlambang²

¹,²Faculty of Engineering and Informatics, Universitas PGRI Semarang, Jl. Sidodadi-Timur No.24 Semarang, Central Java 50232, Indonesia
aristrijaka@upgris.ac.id

Abstract. Sentiment Analysis is now very important and very useful in machine learning technology where a contextual mining of text to identify and extract subjective information in the source, and in helping to understand social sentiment from comments. In general, sentiment analysis can be classified into three broad categories namely sentiment positive and negative. One method of machine learning is the Deep Belief Network (DBN). DBN which is included in the Deep Learning method, is by stacking several algorithms with several extraction features that utilize all resources optimally. This research has two points. First, it aims to classify positive, negative, and neutral sentiments for the test data. The following experiments provide a system of sentiment analysis through the naive Bayes algorithm to calculate sentiment and to improve accuracy by reducing noise in words applied in Indonesian language. From this research, a good level of accuracy can be obtained for extending sentiment using 10-Cross Validation resulting in an accuracy rate of 78.33% with an increase of 6.66% for systems that do not use stopwords, which means reducing noise words and the use of the naïve bayes classification method can be used to determine analysis sentiment on Indonesian language text.

1. Introduction
Nowadays digital data is getting bigger and bigger, and with the current data we are able to determine several things related to the acceleration of decision making, both in the field of trade, government, health and education. According to the internetwordstats survey 4,241,972,790 est. population for Asia in 2019 - Area: 39,365,000 sq km 2,275,469,859 Internet users and 53.6% penetration rate as of July 31/19

| Country     | INDONESIA |
|-------------|------------|
| population  | 269,536,482|
| Area        | 1,904,443 sq km |
| Capital City| Jakarta - 9,751,937 Pop. |
| GNI (per capita) | $3,540 (2017) |
| Internet users in June/2019 | 171,260,000 |
| penetration | 63.5%      |

Based on the data above, the number of internet users is very large, and one of the interactions between internet users is in the form of comments written in a post, both written and video. Comments written in a post are representations of users who interact with the owner to show sentiment. So that an
extraction from the post in the form of tadai text is very much needed considering the amount of data or the number of comments and limitations on the ability to read humans, this requires a machine learning system capable of extracting sentiments in the comment text in Indonesian language.

2. Research Methods

2.1. Process Documents
Tokenization of words(1), We use Tokenization of words to convert sentences into words, the output of this tokenization is words that can be converted into data frames that make it easier for a learning machine to process understanding of the meaning of a word. Tokenization of words is a crucial step in turning words into numerical data for the next process. Transform case(2) the use of transform case is to change all attributes to be the same and the same type so that it will be very easy when doing matching in a text search without constraints of case type differences either uppercase or lowercase. In this case, we change the case to the lowercase for all processed text. Filter Stopwords(3) to delete words that are not needed or unwanted in this project is to filter words using the stopwords library filter, where with a library we can add and subtract words that are not needed in a way that is more flexible and can be placed in a database separate from the system.

2.2. Cross Validation
Cross Validation is a nested operator. Here the operator has two sub processes: Training subprocess and Testing subprocess. The training subprocesses are used to train the system with data that has been labeled. The trained model is then applied to the Testing subprocess. Model performance is measured during the ExampleSet Input Testing phase partitioned into a set of parts k of the same size. From the subset, one subset is retained as a test data set (ie input from the Testing subprocess). Evaluation of the performance of the model on an independent test set results in a good estimate of performance on an invisible data set. This also shows if 'overfitting' occurs. This means that the model represents test data very well, but does not generalize very well to new data.

2.3. Training with Naïve Bayes
In general, the Bayes theorem is stated as follows:

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]  \hspace{1cm} (1)

In this notation \( P(A|B) \) means the probability of event A if B occurs and \( P(B|A) \) chance of event B if A occurs.

2.4. Testing
Cross-validation is one of the statistical methods used to evaluate the performance of models or algorithms where the data is separated into two subsets namely learning process data and validation data. Models or algorithms are trained by a subset of learning and validated by a validation subset. Furthermore the selection of CV types can be based on the size of the dataset. Usually the K-fold CV is used because it can reduce computing time while maintaining the accuracy of the estimate.
The 10-fold CV is one of the recommended K-fold CVs for the selection of the best model because it tends to provide an estimate of accuracy that is less biased compared to the usual CV, leave-one-out CV and bootstrap. In 10 fold CVs, the data is divided into 10 folds of approximately the same size, so we have 10 data subsets to evaluate the performance of the model or algorithm. For each of the 10 data subsets

3. Result
3.1. Read the document
Read the document from comma separated value text to get all documents and then set the role operator functioned in arranging parts of the document that can read and group data into columns, so that it can be identified which columns will be processed as what, for example columns id is an integer which means that the column only contains the index of the data, then the column label contains the binominal type data which is the label of each data record, and the text column contains sentiment data in the form of polynominal type text clusters, so all roles must be determined in accordance with its function correctly in order to obtain the right data.

Each process is connected and can be seen in the picture above, all processes are interconnected with one another in terms of the output of an operator being input for the next operator, and testing of sample documents that are entered and interacted into the system that has been trained to be in check into the operating system.
In Figure 3 is a preprocessing text system where the incoming text is carried out in the initial management by applying the Tokenization operation and then transformed to the lower case and then by filtering the stopword dictionary operator by reducing unnecessary words. The results of preprocessing this text will be input for further processing.

The application of Naïve Bayes algorithm when training data is to be able to provide data to determine the type of sentiment in accordance with the results of the training, and testing by applying a model that has been designed and applied from the start. By measuring performance to get accuracy through 10-Cross validation schemas as a result. The results obtained from this system are as can be seen in the table confusion matrix below.

|           | True negative | True positive | Class precision |
|-----------|---------------|---------------|-----------------|
| Pred. negative | 24            | 11            | 68.57%          |
| Pred. positive  | 6             | 19            | 76.00%          |
| Class recall   | 80.00%        | 63.33%        |                 |

Based Confusion Matrix on table 2 it can be concluded while the resulting vector performave is accuracy: accuracy: 71.67% +/- 19.33% (micro average: 71.67%)
Based on table 3 it can be concluded while the resulting vector performed is accuracy: 78.33% +/- 20.86% (micro average: 78.33%).

4. Conclusion

Based on sentiment testing experiments in Indonesian using the naïve Bayes algorithm, it can be seen in the confusion matrix table above that there are two results, with stopword and without stopword, from the result without stopword yields an accuracy of 71.67% while with stopword dictionary produces an accuracy of 78.33% in this increased by 6.66% which means a significant increase. Technically it can be concluded that word reduction can improve accuracy in sentiment extraction in an Indonesian language text.

Reference lists

[1] M. Barmer, Principles of data mining. 2007.
[2] J. Jotheeswaran, R. Loganathan, and M. S. B, “Feature Reduction using Principal Component Analysis for Opinion Mining,” vol. 3, no. 5, pp. 118–121, 2012.
[3] C. Rohrdantz, M. C. Hao, U. Dayal, L. Haug, H. P. Labs, and D. A. Keim, “Feature-Based Visual Sentiment Analysis of Text Document Streams,” vol. 3, no. 2, 2012.
[4] W. Zhang and F. Gao, “An Improvement to Naive Bayes for Text Classification,” Procedia Eng., vol. 15, pp. 2160–2164, Jan. 2011.
[5] H. Jeong, D. Shin, and J. Choi, “FEROM: Feature Extraction and Refinement for Opinion Mining,” ETRI J., vol. 33, no. 5, pp. 720–730, Oct. 2011.
[6] M. Yusuf and D. D. Santika, “ANALISIS SENTIMEN PADA DOKUMEN BERBAHASA INDONESIA DENGAN PENDEKATAN SUPPORT VECTOR MACHINE,” 2011.
[7] B. Liu, “Sentiment Analysis: A Multi-Faceted Problem,” no. 1, 2010.
[8] V. Roth and T. Lange, “Feature Selection in Clustering Problems,” 2000.
[9] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” Proc. ACL Student Res. Work., 2005.
[10] F. Sebastiani, “Machine learning in automated text categorization,” ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, Mar. 2002.
[11] A. P. Windarto et al., “Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province,” J. Phys. Conf. Ser., vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.
[12] Budiharjo, T. Soemartono, A. P. Windarto, and T. Herawan, “Predicting tuition fee payment problem using backpropagation neural network model,” Int. J. Adv. Sci. Technol., vol. 120, pp. 85–96, 2018, doi: 10.14257/ijast.2018.120.07.
[13] Sumijan, A. P. Windarto, A. Muhammad, and Budiharjo, “Implementation of Neural Networks in Predicting the Understanding Level of Students Subject,” Int. J. Softw. Eng. Its Appl., vol. 10, no. 10, pp. 189–204, 2016.
[14] Budiharjo, T. Soemartono, A. P. Windarto, and T. Herawan, “Predicting School Participation in Indonesia using Back-Propagation Algorithm Model,” Int. J. Control Autom., vol. 11, no. 11, pp. 57–68, 2018.