Classification of Movie Review Sentiment Analysis Using Chi-Square and Multinomial Naïve Bayes with Adaptive Boosting

Muhamad Biki Hamzah [1*]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: biki.hamzah@gmail.com

ARTICLE INFO

Article history
Received 15 February 2021
Revised 18 March 2021
Accepted 12 April 2021

Keywords
Sentiment Analysis
Classification
Multinomial Naïve Bayes
Chi-Square
Adaptive Boosting
Term Frequency-Inverse Document Frequency

ABSTRACT

Sentiment analysis problems have attracted the attention of researchers. Sentiment analysis is a process that aims to determine the sentiment polarity of text. Nowadays, sentiment from product reviews has become a piece of important information for producers and potential customers. This paper conducted a sentiment analysis classification on a movie review from the IMDb site. In the classification analysis, the sentiment of movie reviews used the multinomial naïve Bayes algorithm. Adaboost was applied to boosting the accuracy of multinomial naïve Bayes. Feature selection is used to reduce the number of features and irrelevant features. The chi-square feature selection used was employed in the current study. The accuracy obtained in movie review sentiment analysis classification using the multinomial naïve Bayes algorithm is 81.39%. Meanwhile, the accuracy of the multinomial naïve Bayes algorithm by applying chi-square is 85.37%. The final result of multinomial naïve Bayes algorithm accuracy by applying AdaBoost and chi-square feature selection is 87.74%.

1 Introduction

Sentiment analysis problems, especially classification problems, have attracted the attention of researchers. This problem is also a concern for business actors, both individually and in groups, in marketing their products or services. One of the important supporting factors for prospective buyers to determine the decision to buy a product or service is reviews of the product or service (Zhu & Zhang, 2010).

Sentiment analysis can be defined as the process of finding opinions that have positive, negative, and neutral polarities in data available on social media or websites (Rani & Singh, 2021). According to Bing Liu (2015), sentiment analysis is a field of computational study of opinions, sentiments, and emotions expressed through text. Sentiment analysis can provide benefits in the form of valuable information for various aspects, ranging from service satisfaction levels (Rofiqoh, Perdana, & Fauzi, 2017), election predictions (Sharma & Moh, 2016), and product reviews (Gunawan, Pratiwi, & Pratama, 2018). Currently, consumer opinion is one of the important pieces of information in developing various further products (Jin, Liu, Ji, & Kwong, 2019), one of which is the movie. Consumer opinions can be obtained from blogs, microblogs, discussion forums, and social media. One of the well-known movie database websites is the Internet Movie Database (IMDb). IMDb consists of information about movies and movie production.

Previous researchers have proposed techniques for analyzing product reviews. The proposed methods include naïve Bayes support vector machine (Jagdale, Shirsat, & Deshmukh, 2019), decision tree, logistic regression, random forest, and stochastic gradient descent (Haque, Saber, &

Shah, 2018). Naïve Bayes was chosen in this study. This choice is based on the fact that naïve Bayes is an algorithm that is fast, easy to implement, and effective and useful in high-dimensional data because the probability of each feature is estimated to be independent (Taheri & Mammadov, 2013). Jagdale *et al*. comparing the naïve Bayes algorithm and SVM on different datasets. The study results show that the navenaïve Bayes algorithm has better accuracy on several datasets than the SVM algorithm.

Similar research was also conducted by Baik, Gupta, and Chaplot (2017). Baik *et al*. compare naïve Bayes algorithm, k-nearest neighbour, and random forest on sentiment analysis of movie reviews. The results show that the naïve Bayes algorithm has better accuracy than the k-nearest neighbour and random forest algorithms.

One of the problems that often occur in sentiment analysis is the number of features. A large number of features can reduce the classification performance; therefore, a feature selection process is needed. One of the most popular selection features is the chi-square. Chi-square is the most effective selection feature (Madasu & Elango, 2020). Research conducted by Madasu and Elango on the Amazon Review Dataset, IMDb Review Dataset, and Yelp Review Dataset using several selections features such as odds ratio, chi-square, GSS coefficient, Bi-Normal Separation with logistic regression algorithm, SVM-RBF, SVM -Linear, decision tree, multinomial naïve Bayes, and Bernoulli naïve Bayes. This study shows that multinomial naïve Bayes with chi-square feature selection shows the best accuracy values in two datasets: the Amazon Review Dataset and the IMDb Review Dataset.

In sentiment analysis, to increase accuracy and reduce bias, additional methods are needed. One of the ensemble methods is adaptive boosting (AdaBoost). AdaBoost has the concept of each classifier focused on the previous classifier error by applying the majority voting concept to increase accuracy (Zhang & Ma, 2012). Silva *et al*. (2015) compared the multinomial naïve Bayes algorithm, SVM, SVM with AdaBoost, and multinomial naïve Bayes with AdaBoost. This study shows that multinomial naïve Bayes with AdaBoost has the highest accuracy score of 57.35%.

Based on the description of the problem above, a study was conducted to classify the movie review dataset using the multinomial naïve Bayes algorithm with AdaBoost and the application of the chi-square selection feature.

2 Methods

This study classified the movie review dataset using an approximation method to get better accuracy results. The sentiment analysis classification process was carried out according to Figure 1. The process consists of several main stages, such as the preprocessing stage, data transformation, feature selection, and classification stage.

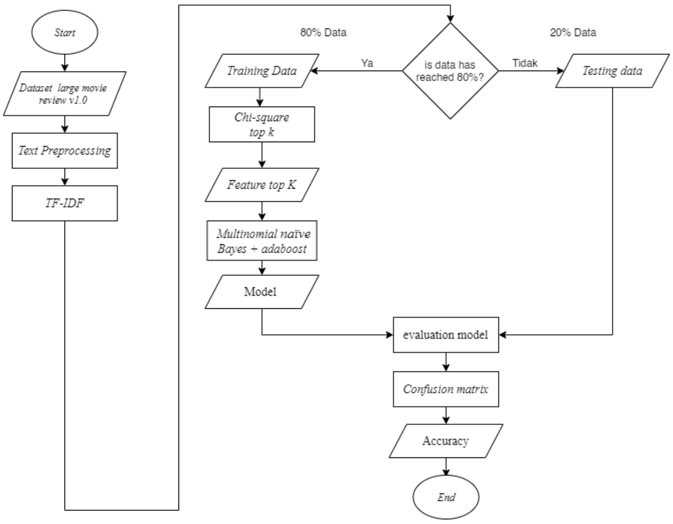

Figure 1. Multinomial naïve bayes algorithm with adaboost and chi-square

2.1 Dataset

The dataset used in this study is the large movie review v1.0 (Maas *et al.*, 2011). dataset obtained from https://ai.stanford.edu/~amaas/data/sentiment. The dataset contains 50,000 data consisting of 25,000 positive data reviews and 25,000 negative data reviews.

2.2 Data Preprocessing

Data preprocessing is the stage of converting unstructured data into structured or semi-structured data [16]. Data preprocessing comprises the following process:

1. Case folding is the process of changing capital letters into non-capital letters (Kowsari *et al.*, 2019).
2. Tokenize is a process to break a document or text into a single word (Susilowati, Sabariah, & Gozali, 2015).
3. The Stopword filter removes words that do not have a contribution to the text classification (Handayani & Pribadi, 2015).
4. Stemming is the process of changing the form of words into basic words (Ipmawati, Kusrini, & Taufiq, 2017).

2.3 Data Transformation

Data transformation is the process of converting tokens into numeric vectors. The purpose of this process is so that the classification algorithm can process the data because the classification algorithm cannot process the original dataset. The method used in this process uses the TF-IDF method. TF-IDF calculates the inverse document frequency (IDF) and frequency term (TF) values for each word or term in each document in a class (Jindal, Malhotra, & Jain, 2015). To calculate TF by counting every word frequency that appears in the document. While the IDF shows the weight of scarcity. The equation for calculating IDF can be seen below.

$$idf(w, D) = \log\left(\frac{N}{f(w,D)}\right) \tag{1}$$

$$tfidf(w, d, D) = tf * idf \tag{2}$$

2.4 Splitting Data

Splitting data is the process of dividing the dataset into training data and test data. The data is divided into 80% training data and 20% test data or 40,000 training data and 10,000 test data. The ratio of the division of 80:20 is chosen based on the Pareto principle (Dunford, Su, Tamang, & Wintour, 2014).

2.5 Chi-Square Feature Selection

Chi-square is the method chosen in this study to reduce the number of features. The selected feature is a feature that has a strong correlation in the classification process. The chi-square equation can be seen in equation 3.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{3}$$

2.6 Multinomial Naïve Bayes

At this stage, the multinomial naïve Bayes algorithm is used to classify the movie dataset reviews that have previously gone through the preprocessing process, data transformation, and selected features. First, calculate the maximum likelihood value. For each term in each class, the equation can be seen in equation 4.

$$\widehat{\mathrm{P}}(\mathrm{t}_k|\mathrm{c}) = \frac{T_{ct} + \alpha}{\left(\sum_{t' \in V} T_{ct'}\right) + B'} \tag{4}$$

After getting the maximum likelihood value, the following process calculates the Prior probability value for each class. The prior probability equation can be seen in equation 5.

$$\widehat{\mathrm{P}}(c) = \frac{N_c}{N} \tag{5}$$

Then calculate the maximum a posteriori value for each class. After that, select the class with the maximum a posteriori value which has the highest value as the class prediction. The maximum a posteriori equation can be seen in equation 6.

$$c_{map} = \arg\max[\log \widehat{\mathrm{P}}(c) + \textstyle\sum_{1\le k\le n} \log \widehat{\mathrm{P}}(\mathrm{t}_k|\mathrm{c})] \tag{7}$$

2.7 Adaboost

Adaboost is an ensemble algorithm that serves to increase the accuracy of classifiers. AdaBoost can be used to reduce errors from base learners consistently to produce better classifications than random guesses (Freund, Schapire, & Hill, 1996). The steps of the AdaBoost algorithm include the following.

1. Initialize the observation weights w_i =1/n, i=1,2,3,…n
2. For $m = 1$ to M:
   a. Fit a classifier $T^{(m)}(x)$ to the training data using weights w_i.
   b. Obtain the weighted class probability estimates

$$p_k^{(m)}(x) = Prob_w(c = k|x), k = 1, \ldots, K. \tag{8}$$

   c. Set

$$h_k^{(m)}(x) \leftarrow (K-1)\left(\log p_k^{(m)}(x) - \tfrac{1}{K}\textstyle\sum_{k'} \log p_k^{(m)}(x)\right), k = 1, \ldots K \tag{9}$$

   d. Set

$$w_i \leftarrow w_i.\exp\left(-\tfrac{K-1}{K} y_i^T \log p^{(m)}(x_i)\right), k = 1, \ldots., K \tag{10}$$

3. Output

$$C(x) = arg\ \max_k \textstyle\sum_{m=1}^{M} h_k^{(m)}(x) \tag{11}$$

3 Results and Discussion

This study applied chi-square feature selection and multinomial naïve Bayes algorithm with AdaBoost to improve sentiment analysis of movie reviews. With AdaBoost and implemented the chi-square selection feature. The higher the accuracy value, the better the algorithm.

In the application of the multinomial naïve Bayes algorithm with chi-square feature selection, it is optimally selected with the highest accuracy value. The results of the optimal feature selection are shown in Figure 2.

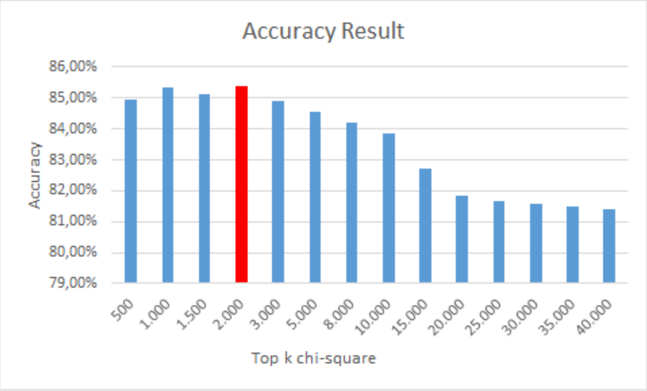

Figure 2. Accuracy result from the experiment of top k chi-square

In the classification stage of the multinomial naïve Bayes algorithm with chi-square feature selection, the optimal number of features is the 2,000 best features. The best number of features will combine with AdaBoost. A fine-tuning process applies the multinomial naïve Bayes algorithm with AdaBoost and chi-square feature selection with the number of selected features 2,000 the best learning rate and iteration values. The results of fine-tuning can be seen in Table 1.

Table 1. Accuracy result (%) from the experiment of AdaBoost

Learning Rate (α)	Iteration				
	50	75	100	200	400
0,1	87,29	87,62	**87,74**	87,53	87,27
0,3	85,57	87,50	87,38	87,22	87,12
0,5	87,50	87,30	87,18	87,15	87,09
0,8	87,29	87,20	87,16	87,09	87,09
1,0	87,19	87,19	87,15	87,10	87,10

The results of the fine-tuning process obtained the highest accuracy value at 100 iterations and a learning rate of 0.1 with an accuracy value of 87.74%. The results of the comparison of the three methods can be seen in Figure 3.

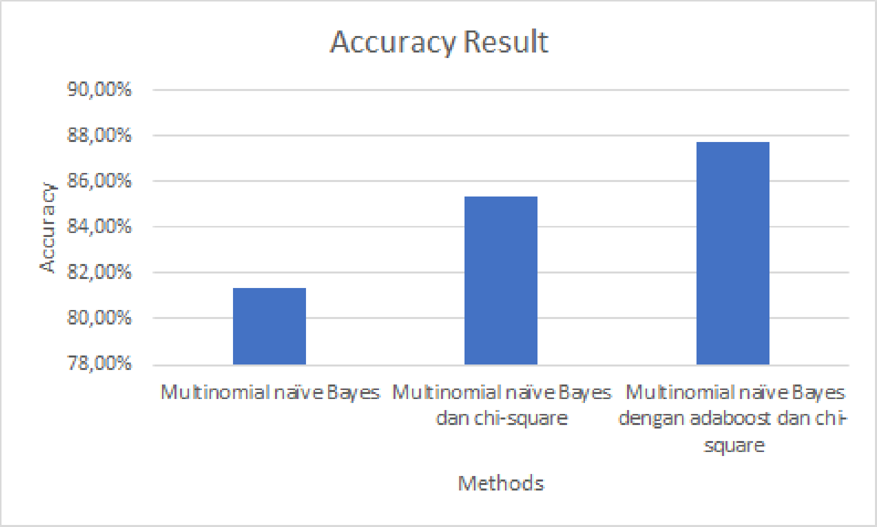

Figure 3. Results of the proposed method

Based on the accuracy values obtained from each algorithm, the accuracy results of applying multinomial naïve Bayes with AdaBoost and chi-square selection features is 6.35%. The application of the multinomial naïve Bayes algorithm with AdaBoost and chi-square feature selection is proven to be able to increase the accuracy of the extensive movie review v1.0 dataset.

4 Conclusion

This paper examines the multinomial naïve Bayes algorithm with AdaBoost to classify sentiment towards film reviews using chi-square. The application of chi-square, in this case, is used to select the best features in the classification process. Adaboost is used to improve the accuracy of the multinomial naïve Bayes algorithm. The evaluation results of the accuracy value show that the combination approach of the multinomial naïve Bayes with AdaBoost and chi-square feature selection produced higher accuracy compared to using only the multinomial naïve Bayes algorithm.

References

Baik, P., Gupta, A., & Chaplot, N. (2017). Sentiment Analysis of Movie Reviews using Machine Learning Classifiers. *International Journal of Computer Applications*, 7(50), 45–49. doi: 10.5120/ijca2019918756

Dunford, R., Su, Q., Tamang, E., & Wintour, A. (2014). The Pareto Principle. *The Plymouth Student Scientist*, 07(1), 140–148.

Freund, Y., Schapire, R. E., & Hill, M. (1996). Experiments with a New Boosting Algorithm Rooms f 2B-428 , 2A-424 g.

Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naïve Bayes [Sentiment Analysis System on Product Reviews Using Naïve Bayes Method]. *Jurnal Edukasi Dan Penelitian Informatika*. doi: 10.26418/jp.v4i2.27526

Handayani, F., & Pribadi, S. (2015). Implementasi Algoritma Naïve Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110 [Implementation of the Naïve Bayes Classifier Algorithm in Automatic Text Classification of Public Complaints and Reporting through Call Center Services 110]. *Jurnal Teknik Elektro*, 7(1), 19–24. doi: 10.15294/jte.v7i1.8585

Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment Analysis on Large Scale Amazon Product Reviews. 2018 IEEE International Conference on Innovative Research and Development (ICIRD 2018), 1–6. doi: 10.1109/ICIRD.2018.8376299

Ipmawati, J., Kusrini, & Taufiq Luthfi, E. (2017). Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen [Comparison of Mining Text Classification Techniques in Sentiment Analysis]. *Indonesian Journal on Networking and Security*, 6(1), 28–36.

Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). *Sentiment Analysis on Product Reviews using Machine Learning Techniques*. In Advances in Intelligent Systems and Computing, 768. Springer: Singapore. doi: 10.1007/978-981-13-0617-4_61

Jin, J., Liu, Y., Ji, P., & Kwong, C. K. (2019). Review on Recent Advances in Information Mining from Big Consumer Opinion Data for Product Design. *Journal of Computing and Information Science in Engineering*, 19(1).

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for Text Classification: Literature Review and Current Trends. *Webology*, 12(2), 1–28.

Kadhim, A. I. (2018). An Evaluation of Preprocessing Techniques for Text Classification. *International Journal of Computer Science and Information Security*, 16(6), 22–32.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information (Switzerland), 10(4), 1–68. https://doi.org/10.3390/info10040150.

Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Morgan & Claypool Publishers. doi: 10.2200/S00416ED1V01Y201204HLT016

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, 3, 23–32.

Madasu, A., & Elango, S. (2020). Efficient feature selection techniques for sentiment analysis. Multimedia Tools and Applications, 79(9–10), 6313–6335. https://doi.org/10.1007/s11042-019-08409-z

Rani, M., & Singh, J. (2021). *A Primer on Opinion Mining: The Growing Research Area*. Advances in Intelligent Systems and Computing, 1165. Springer: Singapore.

Rofiqoh, U., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature [Sentiment Analysis of User Satisfaction Level of Indonesian Cellular Telecommunication Service Providers on Twitter Using Support Vector Machine and Lexion Based Feature Methods]. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, 1(12), 1725–1732.

Sharma, P., & Moh, T. S. (2016). Prediction of Indian election using sentiment analysis on Hindi Twitter. *2016 IEEE International Conference on Big Data*.

Silva, N., Hruschka, E., & Hruschka, E. (2015). Biocom Usp: Tweet Sentiment Analysis with Adaptive Boosting Ensemble. *SemEval*, 129–134. doi: 10.3115/v1/s14-2018

Susilowati, E., Sabariah, M. K., & Gozali, A. A. (2015). Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter [Implementation of the Support Vector Machine Method to Classify Traffic Congestion on Twitter]. *E-Proceeding of Engineering*, 2(1), 1478–1484.

Taheri, S., & Mammadov, M. (2013). Learning the naïve bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, 23(4), 787–795. doi: 10.2478/amcs-2013-0059

Zhang, C., & Ma, Y. (2012). Ensemble Machine Learning Methods and Application. *Journal of Chemical Information and Modeling: Vol. Springer I*.

Zhu, F., & Zhang, X. (2010). Impact of Online Consumer Reviews on Sales:The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing*, 74(3), 133–148.