Comparison of Feature Selection Method in Movie Classification Utilizing Naïve Bayes Classifier

Gita Indah Marthasari[a)], Christian Sri Kusuma Aditya[b)], Muhammad Muzakir Subagio

Department of Informatics, Universitas Muhammadiyah Malang, Malang, Indonesia
Corresponding author: [a)]gita@umm.ac.id
[b)]christianskaditya@umm.ac.id

Abstract. Movies have been currently favored by Indonesians as a means of entertainment medium. Movie fans cover among young adult from all walks of life from young children to the elderly. With the increasing number of fans in Indonesian movies, indirectly encouraging the annual productions of movies. This phenomenon encourages people confusion in watching and preferring about the genre or type of film. Based on this research, it is thus possible to east the problem. This effort assists in grouping or classifying a film. To perform this classification, this study utilizes the Naïve Bayes classification method combining with the selection feature information and the chi square. In this case, the selection of the user features will be compared, which is useful to obtain the best classification value. The results of the classification of films based on synopsis that has the highest results utilize the Naïve Bayes classifier and feature selection of chi square with an accuracy of 90%. The average value of precision is 89%, and the average value of recall is 88%.

INTRODUCTION

In machine learning, text classification serves as the fundamental task and the foundation in natural language preprocessing [1,2,3]. An example from text classification includes Naïve Bayes classifier, an algorithm frequently utilized in classification problem, by constructing the example of training with label class [4]. For the increase of the result using Naïve Bayes classifier to obtain the feature selection. The major characteristic of text classification lies in the number of features in the feature space (vector space, bag of words) straightforwardly reaching the orders of tens of thousands even for moderate size of dataset. Feature selection involves the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Feature selection used for selection of features is contained in dataset prior to performing the classification [5]. Some predictive modeling problems contain a large number of variables hindering the development and training of models and requiring a large amount of system memory. Additionally, the performance of some models could degrade when including input variables that are irrelevant to the target variable. The utilized feature selection this research includes the information and chi square. Besides, other features of selection algorithm are commonly employed, such as: particle swarm optimization, document frequency thresholding, genetic algorithm, etc. [6].

At this time, films have gained popularity by Indonesians. Movie fans cover not only among young adult but from all walks of life from young children even the elderly. Along with the growing number of fans in Indonesian movies, this phenomenon leads to the growing production of movies annually. However, this also drives people's confusion in terms of selecting movie genre. Other than that, movie gains popularity due to the rapid development of the internet [7], where information is limitlessly accessed.

Several studies have been conducted prior to movie classification. Previous study [8], utilized the improved k-nearest neighbor. In [2], the three algorithms were utilized to compare methods such as: support vector machine, multinominal naïve bayes and multilayer perceptron and mixed with the bag of words and tf-idf. In [9], the four algorithms were utilized to compare decision tree, random forest, extra trees and multi-layer perceptron and mixed all of the algorithms with tf-idf, word2vec, fastText, wang2vec, glove and doc2vec. In [7], researchers classify the movie by employing method support vector machine along with classifying the poster and the synopsis of the movies. In [10], researcher utilized neural network architecture to classify movies based on the textual synopsis.

Based on previous research, the focus in this paper is to classify movie based on the textual synopsis applying Naïve Bayes classifier and to compare feature selection information and chi square. Classifier would further be evaluated with used precision, recall and accuracy.

METHOD

The subject in this research is to compare the feature selection information and chi square implemented with Naïve Bayes classifier to classify film based on synopsis of the film. The dataset consists of 4 classes with a

1st International Conference on Technology, Informatics, and Engineering
AIP Conf. Proc. 2453, 030006-1–030006-6; https://doi.org/10.1063/5.0094436
Published by AIP Publishing. 978-0-7354-4356-3/$30.00

balanced number of distributions as illustrated in Table 1. The dataset is further distributed in seven conditions for each feature selection to obtain the best value of evaluate.

TABLE 1. Class distribution

Movie Class	Number of Synopsis
Horror	100
Drama	100
Comedy	100
Action	100

Information Gain

Information gain includes the feature selection to calculate a value of each feature and selected features that are not correlated with the information of the data would be removed [11]. The removed features are measured by calculating the entropy of each feature and corps analysis [12][13]. As a result, the obtained information is utilized as features that are essential to a category [14], where higher value of information indicates more important characteristics [15].

The calculation of the information is performed by Equation 1 to obtain the value of information on each word in dataset.

$$InfoGain\ (S, A) = Entropy(S) - \sum_{v} \in Value(A) \frac{|S_v|}{s} Entropy(S_v) \quad (1)$$

The calculation of the entropy is performed by Equation 2 to obtain the score of entropy on each word.

$$Entropy(S) = -\sum \frac{|S_i|}{s} log \frac{S_i}{s} \quad (2)$$

Upon obtaining the score of information, the features having smallest value which are contained in dataset would be deleted.

Chi Square

Chi square contains the feature selection algorithm performed to select the range of irrelevant features in classification process [16]. The chi square is utilized to improve a classification performance by removing noisy data and selecting a represent subset of the dataset to reduce intricacy of the classification [17]. Chi square is additionally employed to measure the relationship between class and feature to be compared with one degree of freedom [18]. If the value of feature are high, the relationship between feature and class is high [19].

The calculation of chi square value is conducted by Equation 3 to obtain the value of chi square on each word in dataset.

$$X^2(t, c) = \frac{N(A \times D - B \times C)^2}{(A + B) \times (C + D) \times (A + C) \times (B + D)} \quad (3)$$

Naïve Bayes Classifier

Naïve bayes becomes one of classification algorithms applied in theorem bayes [20]. Naïve bayes classifier provides an efficiency of classification result when implemented in analysis text data such as in document categorization and email spam filtering [21]. The algorithm performed the classification of probabilities and statistic by predicting opportunities in the future based on past experience [22][23]. The bayes rules provide a foundation about the algorithm that defines formula containing a relationship among dependent variables including future and past experience [24]. Naïve bayes provides assumptions that the existence of certain features in a category or class is not related with any other feature [25].

The calculation of document C to class classification is measured by Equation 4 to obtain the score of probability of document C.

$$P(C|F_1, \dots, F_n) = P(C)\ P(F_1|C)P(F_2|C)\ P(F_3|C) \dots\ P(F_n|C) \quad (4)$$

The calculation for the probability of category C is measured by Equation 5 to obtain the score of probability with similar category.

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

The calculation for the probability of word from category C is measured by Equation 6 to obtain the score of probability from word with similar category.

$$P(F_n|C) = \frac{count(tn, c) + 1}{count\ (c) + |V|} \tag{6}$$

Scenario

Testing scenario aims to obtain the highest value from accuracy, precision and recall in this research. Testing scenario that uses in this research is conducted by supply training test to test the classification by distributing dataset into training data and testing data. Data will be distributed into seven conditions. First, data is distributed into 25% training data and 75% testing data. Second, data is distributed into 50% training data and 50% testing data. Third, data is distributed into 75% training data and 25% testing data. Fourth, data is distributed into 80% training data and 20% testing data. Fifth, data is distributed into 85% training data and 15% testing data. Six, data is distributed into 90% training data and 10% testing data. The last, data is distributed into 95% training data and 5% testing data.

RESULT

This research attempts to compare feature selection information gain and chi square by utilizing Naïve Bayes classifier. During the classification, data is distributed as described in testing scenario. As aforementioned, classification is evaluated with evaluation models of precision, recall and accuracy.

Information Gain

First, the textual synopsis of the film is extracted by utilizing information gained in dataset. From that the utilized feature selection, the 5796 features are obtained in dataset, leaving 2482 features excluded for classification, which has no effect with the classifier indicated by the low value of information. In the classification that employed Naïve Bayes classifier, presenting the seven conditions of distributed dataset, illustrated in Table 2.

TABLE 2. Comparison of distribute scenario Information Gain result

Scenario	Precision	Recall	Accuracy
25% Training and 75% Test	62%	63%	61%
50% Training and 50% Test	63%	63%	65%
75% Training and 25% Test	70%	71%	70%
80% Training and 25% Test	73%	72%	74%
85% Training and 15% Test	69%	70%	67%
90% Training and 10% Test	77%	77%	75%
95% Training and 5% Test	86%	88%	85%

Results from the classification employed in distributing dataset in scenario 7 obtain the average results of recall of 88%, average of precision of 86% and classification accuracy of 85%. Upon implementing the distributed data, classification obtained the best value of evaluation compared with other distributed data by utilizing feature selection information. The result of the classification is illustrated on Table 3.

TABLE 3. The result of best classification on feature selection Information Gain

Real	Class				Total Class
	Drama	Horror	Comedy	Action	
Drama	6	0	1	0	7
Horror	0	4	0	0	4
Comedy	2	0	3	0	5
Action	0	0	0	0	4

Chi Square

Lastly, the feature selection of chi square is utilized to extract the textual synopsis of movies from dataset. The obtained features involve 1598 out of 5715 features in dataset. Alike in the previous experiment, the classification used naïve bayes classifier generating the seven distributed data in classification, illustrated in Table 4.

TABLE 4. Comparison of Chi Square result

Scenario	Precision	Recall	Accuracy
25% Training and 75% Test	56%	55%	56%
50% Training and 50% Test	66%	67%	66%
75% Training and 25% Test	70%	70%	72%
80% Training and 25% Test	70%	72%	68%
85% Training and 15% Test	70%	68%	72%
90% Training and 10% Test	76%	77%	75%
95% Training and 5% Test	88%	89%	90%

Result from the classification utilizes the distributed dataset in scenario 7 and information gained obtains the average results of recall 89%, precision average of 88% and classification accuracy of 90%. Alike the classification utilizing naïve bayes classifier mixed with feature selection information gain, the distributed data obtains the best value of evaluation compared with other distributed data, illustrated in Table 5 presenting the best classification.

TABLE 5. The result of best classification on Feature Selection Chi Square

Real	Class				Total Class
	Drama	Horror	Comedy	Action	
Drama	6	1	0	0	7
Horror	0	2	0	1	3
Comedy	0	0	2	0	2
Action	0	0	0	8	8

For feature selection of chi square utilizes 1598 features. Meanwhile, the feature selection for information classification utilizes 2482 features, further compared with the feature selection of chi square to reduce more feature than information gain. In the dataset, there was feature that used with both of feature selections or otherwise, as illustrated in Figure 1.

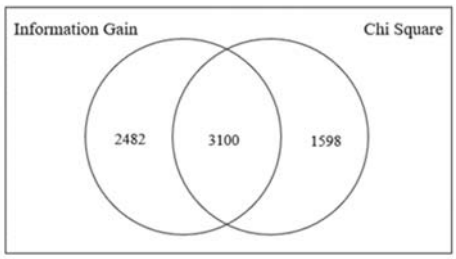

FIGURE 1. Number of features

CONCLUSION

The aim of dimensionality reduction is to reduce vector space and to avoid the overfitting without sacrificing the performance of the classification, overcome by feature selection. Based on this research, the result classification of naïve bayes classifier using feature selection of chi square obtains the best classification. Upon

utilizing the distributed data, including: 95% training data and 55% testing data, 90% accuracy result, 89% average value of recall and 88% average value of precision. In accordance with the classification mixed with feature selection information gain, the distributed data with 95% training data and 5% testing data obtains the best value. Higher testing data present the best result of classification. However, classification mixed with information obtains the accuracy of 85%, recall average value of 88% 88% and precision average value of 86%. It is found that Chi Square presents the most effective result in aggressive term removal without losing classification accuracy in this experiment with naive bayes classifier. Hence, it is concluded that feature selection of chi square is better than feature selection of information.

REFERENCES

1. C. Du, Z. Chen, F. Feng, L. Zhu, T. Gan, and L. Nie, "Explicit Interaction Model towards Text Classification," Proceedings of the AAAI Conference on Artificial Intelligence, (2019).
2. A. C. Saputra, A. B. Sitepu, Stanley, P. W. P. Yohanes Sigit, P. G. Sarto Aji Tetuko, and G. C. Nugroho, "The Classification of the Movie Genre based on Synopsis of the Indonesian Film," Proceeding - 2019 International Conference of Artificial Intelligence and Information Technology, ICAIIT 2019, 201–204 (2019).
3. K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information (Switzerland) (2019).
4. L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," Engineering Applications of Artificial Intelligence (2016).
5. N. K. Verma and A. Salour, "Feature Selection," Intelligent Condition Based Monitoring: For Turbines, Compressors, and Other Rotating Machines, Singapore: Springer Singapore, 175–200 (2020).
6. A. Kumar, R. Khorwal, and S. Chaudhary, "A survey on sentiment analysis using swarm intelligence," Indian Journal of Science and Technology, 9, (39) (2016).
7. X. He et al., "Intelligence science and big data engineering: Image and video data engineering: 5th international conference, IScIDE 2015 Suzhou, China, june 14-16, 2015 revised selected papers, Part I," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), **9242**, 1 (2015).
8. N. Muslimah and R. C. Wihandika, "Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbor (K-NN)," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, **3**, (1), 196–204 (2019).
9. G. Portolese and V. D. Feltrin, "On the Use of Synopsis-based Features for Film Genre Classification," 892–902, 2019.
10. J. Wehrmann, M. A. Lopes, and R. C. Barros, "Self-attention for synopsis-based multi-label movie genre classification," Proceedings of the 31st International Florida Artificial Intelligence Research Society Conference, FLAIRS 2018, 236–241 (2018).
11. M. Y. Abu Bakar, Adiwijaya, and S. Al Faraby, "Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation)Using Information Gain and Backpropagation Neural Network," Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018, 344–350 (2019).
12. I. Kurniawati and H. F. Pardede, "Hybrid Method of Information Gain and Particle Swarm Optimization for Selection of Features of SVM-Based Sentiment Analysis," 2018 International Conference on Information Technology Systems and Innovation (ICITSI), 1–5 (2018).
13. F. He, H. Yang, Y. Miao, and R. Louis, "A Hybrid Feature Selection Method Based on Genetic Algorithm and Information Gain," 320–323 (2016).
14. Y. Liu, X. Yi, R. Chen, Z. Zhai, and J. Gu, "Feature extraction based on information gain and sequential pattern for English question classification," 520–526 (2018).
15. J. Xu and H. Jiang, "An Improved Information Gain Feature Selection Algorithm for SVM Text Classifier," (2015).
16. Y. D. Setiyaningrum, "Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm," International Seminar on Application for Technology of Information and Communication (iSemantic), 1–4 (2019).
17. S. Bahassine, E. Jadida, A. Madani, E. Jadida, and M. Kissi, "An improved Chi-sqaure feature selection for Arabic text classification using decision tree," (2016).
18. A. W. Haryanto and E. K. Mawardi, "Influence of Word Normalization and Chi-squared Feature Selection on Support Vector Machine (SVM) Text Classification," 2018 International Seminar on Application for Technology of Information and Communication, 229–233 (2018).
19. M. Fanbo and C. H. I. Statistics, "An Improved Native Bayes Classifier for Imbalanced Text Categorization Based on K-means and CHI-square Feature Selection," 2018 Eighth International Conference on

Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 894–898 (2018).
20. A. M. Rahat, A. Kahir, A. Kaisar, and M. Masum, “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset,” (2019).
21. D. Buži and J. Dobša, “Lyrics Classification using Naive Bayes,” International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1011–1015 (2018).
22. A. D. Hartanto, “Job Seeker Profile Classification of Twitter Data Using the Naïve Bayes Classifier Algorithm Based on the DISC Method,” 533–536 (2019).
23. R. A. Putri, S. Sendari, and T. Widiyaningtyas, “Classification of Toddler Nutrition Status with Anthropometry Calculation using Naïve Bayes Algorithm,” 2018 International Conference on Sustainable Information Engineering and Technology (SIET), 66–70 (2018).
24. A. O. Adi, “20 Haber Grubu ’ nun Naïve Bayes Yöntemi ile S ı n ı fland ı r ı lmas ı Classification of 20 News Group with Naïve Bayes Classifier,” 2150–2153 (2014).
25. A. Goel and J. Gautam, “Real Time Sentiment Analysis of Tweets Using Naive Bayes,” 257–261 (2016).