Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11788
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The effect of Chi-Square Feature Selection on Question Classification using Multinomial Naïve Bayes

Novi Yusliani[1)], Syechky Al Qodrin Aruda[2)], Mastura Diana Marieska[3)], Danny Mathew Saputra[4)], Abdiansah[5)]
[1,2,3,4,5)]Universitas Sriwijaya, Indonesia
[1)]novi_yusliani@unsri.ac.id, [2)]syechkyal@gmail.com, [3)]mastura.diana@ilkom.unsri.ac.id,
[4)]danny.saputra@gmail.com, [5)]abdiansah@unsri.ac.id

Submitted : Sep 10, 2022 | **Accepted** : Oct 6, 2022 | **Published** : Oct 9, 2022

Abstract: Question classification is one of the essential tasks for question answering system. This task will determine the expected answer type (EAT) of the question given to the system. Multinomial Naïve Bayes algorithm is one of the learning algorithms that can be used to classify questions. At the classification stage, this algorithm used a set of features in the knowledge model. The number of features used can result in curse of dimensionality if the feature is in high dimension. Feature selection can be used to reduce the feature dimension and could increase the system performance. Chi-Square algorithm can be used to select features that describe each category. In this research, the Multinomial Naïve Bayes is used to classify the question sentences and the Chi-Square algorithm is used for the feature selection. The dataset used is a set of Indonesian question sentences, consisting of 519 labeled factoids, 491 labeled non-factoids, and 185 labeled other. The test results showed an increase in accuracy of 0.1 when used feature selection. System accuracy when used feature selection is 0.87 with the number of features used are 248. Without feature selection, the accuracy is 0.77 with the number of features used are 1374.

Keywords: Chi-Square Algorithm; Factoid Question; Feature Selection; Multinomial Naïve Bayes Algorithm; Non-Factoid Question

INTRODUCTION

Question analysis is one of the essential components in question answering system. This component is used to determine the expected answer type (EAT) and question keywords. The expected answer type is regard to category of the question given to the question answering system. In question analysis component, question classification is one of the essential tasks to classify the question into a category. Question classification role is to predict the form of precise response according to the query (Zulqarnain, et.al., 2021). The primary concern in question analysis is to extract useful information from a given question to be used in subsequent modules to finally generate a correct response (Derici, et. al., 2015). There are two approaches for question classification; rule-based, and machine-learning based (Faiz Ur Rahman Khilji, et. al., 2020). Rule-based approach needs to generate many rules when applied in a different contextual setting (Faiz Ur Rahman Khilji, et. al., 2020). To handle this issue, a machine-learning based approach is introduced. In machine-learning based approach, a set of labeled data is used to get a knowledge model. Then, the classifier classifies new data using the knowledge model.

In classification task, the Naïve Bayes algorithm method is effective and performs better than any other method (Abdurrahman Farisi, et.al., 2019). Naïve Bayes algorithm method is statistical classification algorithm based. It used the Bayes' theorem, to find the conditional probability of happening of two events based on the probabilities of happening of each individual event. Two of Naïve Bayes algorithm methods are used for document classification, Multinomial Naïve Bayes algorithm and Bernoulli Naïve Bayes algorithm. The performance of Multinomial Naïve Bayes algorithm performs slightly better than Bernoulli Naïve Bayes algorithm (Singh, et. al., 2019).

Multinomial Naïve Bayes algorithm uses term frequency to describe how many times does the word occur in a document. Term frequency gives fact that whether the word occur in a document or not as well as its frequency in that document (Singh, et. al., 2019). To predict a category of a text, Multinomial Naïve Bayes

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11788
e-ISSN : 2541-2019
p-ISSN : 2541-044X

algorithm multiply the probabilities of the occurrence of all terms in the text against all categories and the one which is higher gives the category to the text. Term used in training phase and testing phase known as feature. Feature dealing with the dimensionality space. Too many features used, will increased the dimensionality space. The curse of dimensionality is a phenomenon that arises when analyzing data in high-dimensional spaces. Issues when dealing data in high-dimensional spaces, such as poor classification accuracy and stalled genetic search (Debie & Shafi, 2019). To solve this issue, feature selection plays a vital role which is modeled to select the feature set from the greater number of features from the high-dimensional data. Feature selection builds the task becomes simple and gives the higher accuracy (Manikandan & Abirami, 2021). By feature selection, it will simplify the model by reducing the number of parameters, next to decrease the training time, to reduce overfilling by enhancing generalization, and to avoid the curse of dimensionality (Chen, et. al., 2020). Statistical based algorithm can be used for feature selection is Chi-Square. In this research, the performance of Indonesian Question classification using Multinomial Naïve Bayes algorithm with Chi-Square algorithm will be compared with the performance of Indonesian Question classification using Multinomial Naïve Bayes algorithm without Chi-Square algorithm.

LITERATURE REVIEW

Question answering system is a system that automatically answers questions in the users' natural language (Hanifah & Kusumaningrum, 2021). This system consists of three main components: question analysis, document retrieval, and answer finder. In question answering system, predicting the entity type of the answering sentence for a given question is the task of question analysis. This task is done by question classification. Question classification plays an important role in finding or constructing accurate answers (Van-Tu & Anh-Cuong, 2016). It can improve the quality of question answering system. In this research, question sentence classifies into three categories. Three categories used in this research are factoid, non-factoid, and other. Classifying a question into a category is supervised learning task in machine learning based. One such algorithm based on supervised learning task is Multinomial Naïve Bayes algorithm.

Multinomial Naïve Bayes algorithm is one of Naïve Bayes Classifiers. It is the statistical classification algorithm based on the Bayes' theorem. Multinomial Naïve Bayes algorithm is widely used in text classification due to its simplicity, efficiency, and efficacy (Jiang, et. al., 2016). This algorithm used set of term known as feature in training phase and in testing phase. Feature dealing with dimensionality space. Too many features used in training phase and testing phase, will increased the dimensionality space. This issue can be solved by using feature selection. Feature selection is an operation of finding out feature words with relatively large effects and removing feature words that have little effects on classification performance (Zai, et. al., 2018). Using feature selection, redundant and irrelevant attributes are removed, then select the most distinct features (Bahassine, et. al., 2020). Feature selection used in this research is Chi-Square algorithm.

Chi-Square algorithm, a statistical based algorithm, is one of the most efficient feature selection methods (Jin, et. al., 2015). In feature selection, there are two variables which refer to the occurrence frequency of feature *t* and the probability of occurrence of category *c* respectively. Chi-Square algorithm conducts a significance test on the relationship between the values of feature and the category (Peker & Kubat, 2021). The Chi-Square formula is related to information-theoretic feature selection functions which try to capture the intuition that the best feature t_k for the category c_i are the ones distributed most differently in the sets of positive and negative examples of class c_i (Bahassine, et. al., 2020). Equation 1 shows the Chi-Square formula.

$$Chi - Square(t_k, c_i) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (1)$$

Where, *N* is total number of documents in the dataset, *A* is number of documents in category c_i that contain the feature t_k, *B* is number of documents that contain the feature t_k in other categories, *C* is number of documents in category c_i that do not contain the feature t_k, D is number of documents that do not contain the feature t_k in other categories.

METHOD

In this research, there are five main components. First component is pre-processing. Pre-processing is a process to remove unnecessary data contained in the text that does not match the required process (Sitepu, et. al., 2022). This component used to represent sentence question into a set of term. This set of term known as feature in classification task. To represent question sentence into feature, question sentence has to passed four processes in pre-processing component. Four processes done in pre-processing component are case folding, tokenization, noise removal, and stemming. Case folding changes all characters in question sentence into small case. By using white space, tokenization splits sentence into an array of word. In this research, filtering is done by using noise removal. The last process in pre-processing component is stemming, making word into its stemmed. These stem

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11788
e-ISSN : 2541-2019
p-ISSN : 2541-044X

words used as feature in classification task. Second component is feature selection. This component used to select feature using Chi-Square algorithm. Then, weighting the feature using *tf-idf*. *tf-idf* is an unsupervised term weighting method. It is the product of the term frequency component (*tf*) by the collection frequency component (*CF*): term frequency (*tf*) and inverse document frequency (*idf*) respectively (Mazyad, et. al., 2017). After feature weighting, training is done to get the knowledge model. To get the knowledge model, this research used Multinomial Naïve Bayes (MNB) algorithm.

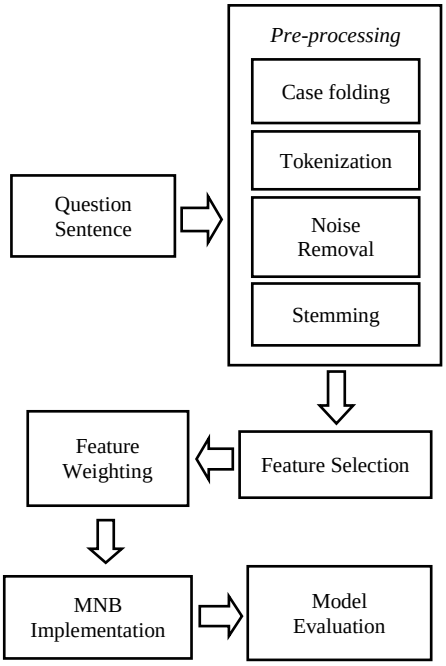

Fig. 1 Architecture System

Multinomial Naïve Bayes (MNB) algorithm is statistics algorithm based using Bayes' theorem (Syahputra, et. al., 2022). This algorithm assumes independence among feature once the category they belong to is known. This model is not only described those features appearing in each document but also the frequency of appearance of each feature. Based on this, a high appearance frequency increases the probability of belonging to a particular category (Bermejo, et. al., 2011). In this research, classification is done by using the highest probability of belonging to a particular category to classify question sentence. Then the last component is model evaluation to evaluate the performance of the system. To evaluate the model, this research used a set of Indonesian question sentence. Fig. 1 describes the architecture of this research.

Table 1. Question Sentence Example for Each Category

Question Sentence	**Category**
What is meant by symbiotic mutualism? (Apa yang dimaksud dengan simbiosis mutualisme?)	*Non-Factoid*
Who was the first President of Indonesia? (Siapa presiden Indonesia yang pertama?)	*Factoid*
Why do plants need sunlight? (Kenapa tumbuhan membutuhkan sinar matahari?)	*Non-Factoid*
When was the tsunami in Aceh happened? (Kapan peristiwa tsunami di Aceh terjadi?)	*Factoid*
Can you translate this paragraph? (Apakah kau bisa menerjemahkan paragraf ini?)	*Other*

Dataset used in this research is a set of Indonesian question sentence labeled with factoid, non-factoid, and other. This dataset consists of 519 question sentences labeled factoid, 491 question sentences labeled non-factoid, and 185 question sentences labeled other. Factoid question is a question requiring just one answer or statement of fact (Herrera, et. al., 2019). Non-factoid question is question used to gain understanding of something. This type of question mostly using question words, such as 'what', 'why', and 'how'. In science, non-factoid questions represent questions regarding definition, reason, and method (Hanifah & Kusumaningrum,

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

2021). Category 'other' in this research is questions that are not in factoid category and non-factoid category, such as list, yes-no, and opinion. Table 1 presents the example of each category.

RESULT

This research used 10-fold cross validation method, which is split dataset into 10. To evaluate the effect of Chi-Square algorithm in question classification using Multinomial Naïve Bayes algorithm, there are two experiment scenarios. Scenario one tested the question classification using Multinomial Naïve Bayes algorithm without feature selection and scenario two tested the question classification using Multinomial Naïve Bayes algorithm to classify the question sentence and Chi-Square algorithm to select the feature. Table 2 shows the performance of question classification using Multinomial Naïve Bayes algorithm based on precision, recall, f-measure, and accuracy value. Number of feature used in scenario one is 1374.

Table 2. Performance Question Classification using Multinomial Naïve Bayes Algorithm

Precision	Recall	F-Measure	Accuracy
0.73	0.73	0.73	0.77

In scenario two, we used nine thresholds to select the feature. Table 3 shows number of features used for each threshold in question classification using Multinomial Naïve Bayes algorithm and Chi-Square algorithm. From table 3, the threshold value determines the number of selected features. In classification, feature set influence the learning process of the classifier. The number of features used has an important role in determine the performance of question classification system. A large number of features used, does not guarantee the question classification system will give the best performance. Fig. 2 shows the accuracy of question classification system for each threshold. Threshold with the highest accuracy is 4.5 using 248 features. Table 4 shows the performance of question classification system based on precision, recall, f-measure, and accuracy value using 4.5 as the threshold.

Table 3. Number of Features for Each Threshold

Threshold	Number of Features
3	318
3.5	309
4	269
4.5	248
5	244
5.5	130
6	118
6.5	109
7	103

Table 4. Performance Question Classification using Multinomial Naïve Bayes Algorithm and Chi-Square Algorithm

Precision	Recall	F-Measure	Accuracy
0.86	0.84	0.84	0.87

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2022
DOI : https://doi.org/10.33395/sinkron.v7i4.11788
e-ISSN : 2541-2019
p-ISSN : 2541-044X

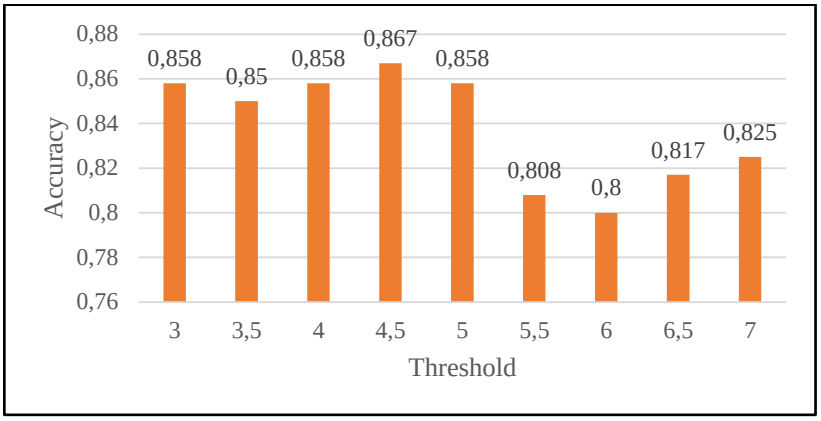

Fig. 2 Accuracy for Each Threshold

DISCUSSIONS

Question classification is a classification task to classify a question sentence. Based on the answers generated, there are five categories, non-factoid question, factoid question, yes-no question, list question, and opinion question (Purwarianti & Yusliani, 2011). Factoid question and non-factoid question are question category usually asked by user in question answering system. Dataset used in this research is a set of question sentence labeled by factoid, non-factoid, and other. In classification, labeled dataset is to be given as training set to the classifier. Multinomial Naïve Bayes algorithm is a classifier algorithm using Bayes' theorem. This algorithm is suitable for text classification since its conditional probability is computed based on the feature vector values (Santhi & Brindha, 2019). The popular issue in classification task is when the feature growth exponentially then the feature vector dimension is high. This issue known as curse of dimensionality.

Feature selection is a task to reduce the dimension of feature vector by selecting the feature. Chi-Square algorithm selects the feature based on the relationship value between feature and category occurrence. If the score is high, feature to category is more likely to be correlated (Cascaro, et. al., 2019). The purpose of this research is to observe the effect of Chi-Square algorithm in question classification system using Multinomial Naïve Bayes algorithm. Fig. 3 shows the performance of question classification using Multinomial Naïve Bayes algorithm with feature selection and without feature selection. It shows that the performance of question classification using Multinomial Naïve Bayes (MNB) algorithm with feature selection using Chi-Square (CS) algorithm is better than using Multinomial Naïve Bayes algorithm without feature selection. Value in each performance measure when using feature selection is increased.

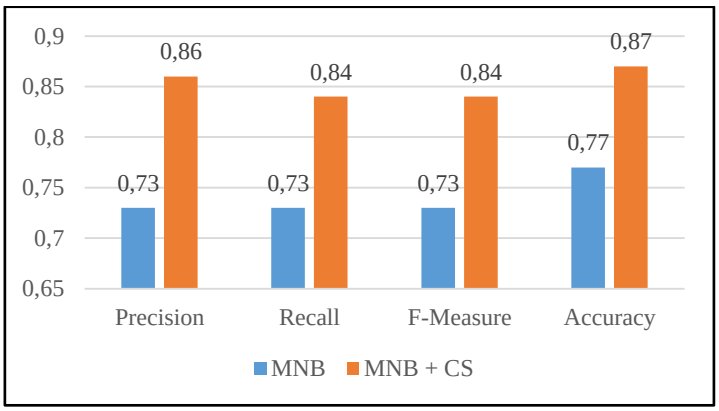

Fig. 3 Performance of Question Classification with Feature Selection and without Feature Selection

Number of features used in question classification system using Multinomial Naïve Bayes algorithm without feature selection is 1374. When using feature selection, number of features used is 248. Features selected are features with the relationship value more than 4.5. Fig. 4 shows the number of features used in each model.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

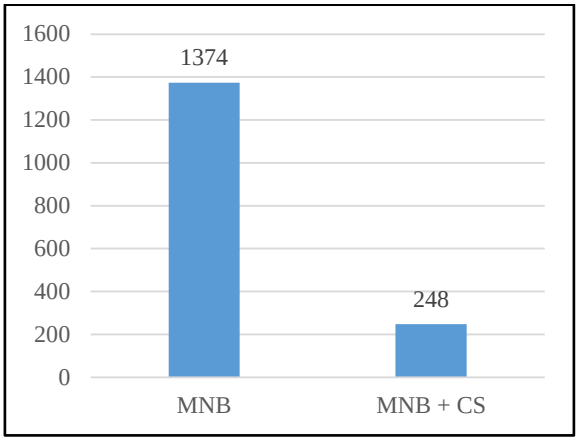

Fig. 4 Number of Features used in Question Classification with Feature Selection and without Feature Selection

CONCLUSION

Feature selection in question classification system using Multinomial Naïve Bayes algorithm improved the performance of the question classification system. Feature selection algorithm used in this research is Chi-Square algorithm. In accuracy, the value is increased 0.1 from 0.77 to 0.87. In other performance measure, such as precision, recall, and f-measure, values are also increased. Number of features used is decreased from 1374 features to 248 features. The experiment results show that Chi-Square algorithm enhanced the performance of question classification system using Multinomial Naïve Bayes algorithm with number of features used is less. For future work, the imbalanced data in dataset issue should be addressed. Also, the performance of other feature selection algorithm could be compared to determine which is the best feature selection algorithm for question classification system.

ACKNOWLEDGMENT

The research of this article was funded by DIPA of Public Service Agency of Universitas Sriwijaya 2021, SP DIPA-023.17.2.677515/2021, on November 23rd, 2020. In accordance with the Rector's Decree Number: 0007/UN9/SK.LP2M.PT/2021, on April 27th, 2021.

REFERENCES

Abdurrahman Farisi, A., Sibaroni, Y., & Al Faraby, S. (2019). Sentiment Analysis on Hotel Reviews using Multinomial Naïve Bayes Classifier. *The 2nd International Conference on Data and Information Science*.

Bahassine, S., Madani, A., Al-Sarem., M., & Kissi, M. (2020). Feature Selection using An Improved Chi-Square for Arabic Text Classification. *Journal of King Saud University – Computer and Information Sciences*, vol.32(2), 225-231. https://doi.org/10.1016/j.jksuci.2018.05.010.

Bermejo, P., Games, J., A., & Puerta, J., M. (2011). Improving The Performance of Naïve Bayes Multinomial in e-mail Foldering by Introducing Distribution-based Balance of Datasets. *Expert Systems with Applications*, vol. 38(3), 2072-2080. https://doi.org/10.1016/j.eswa.2010.07.146.

Cascaro, R. J., Gerardo, B. D., & Medin, R. P. (2019). Aggregating Filter Feature Selection Methods to Enhance Multiclass Text Classification. Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City, 80-84. https://doi.org/10.1145/3377170.3377209.

Chen, R., Dewi, C., Huang, S., & Eko Caraka, R. (2020). Selection Critical Features for Data Classification Based on Machine Learning Methods. *Journal of Big Data*, 7:52. https://doi.org/10.1186/s40537-020-00327-4.

Debie, E., & Shafi, K. (2019). Implications of the Curse of Dimensionality for Supervised Learning Classifier Systems: Theoretical and Empirical Analyses. *Pattern Analysis & Applications*, vol. 22, issue 2, 519-536.

Derici, C., Celik, K., Kutbay, E., Aydin, Y., Gungor, T., Ozgur, A., & Kartal, G. (2015). Question Analysis for A Closed Domain Question Answering System. *Springer International Publishing*. Doi: 10.1007/978-3-319-18117-2_35.

Faiz Ur Rahman Khilji, A., Manna, R., Rahman Laskar, S., Pakray, P., Das, D., Bandyopadhyay, S., & Gelbukh, A. (2020). Question Classification and Answer Extraction for Developing a Cooking QA System. *Computacion y Sistemas*, vol. 24, no.2, 927-933.

Hanifah, A. F., & Kusumaningrum, R. (2021). Non-Factoid Answer Selection in Indonesian Science Question Answering System using Long Short-Term Memory (LSTM). *Procedia Computer Science*, vol.179, 736-746. https://doi.org/10.1016/j.procs.2021.01.062.

*name of corresponding author

Herrera, J., Parra, D., & Poblete, B. (2019). Social QA in Non-CQA Platforms. *Future Generation Computer Science*, vol.105, 631-649. https://doi.org/10.1016/j.future.2019.12.023.

Jiang, L., Wang, S., Li, Z., & Zhang, L. (2016). Structure Extended Multinomial Naïve Bayes. *Information Sciences Journal*, vol. 329, 346-356.

Jin, C., Ma, T., Hou, R., Tang. M., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2015). Chi-Square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization. *IETE Journal of Research*, vol. 61(4), 351-362, https://doi.org/10.1080/03772063.2015.1021385.

Manikandan, G., & Abirami, S. (2021). Feature Selection and Machine Learning Models for High-Dimensional Data: State-of-the-Art. *Computational Intelligence and Healthcare Informatics*, Wiley Online Library.

Mazyad, A., Teytaud, F., & Fonlupt, C. (2017). A Comparative Study on Term Weighting Schemes for Text Classification. *The 3rd International Conference on Machine Learning, Optimization, and Big Data*, Tuscany, Italy, 100-108. Doi: 10.1007/978-3-319-72926-8_9.

Peker, N., & Kubat, C. (2021). Application of Chi-Square Discretization Algorithms to Ensemble Classification Methods. Expert Systems With Applications, vol. 185. https://doi.org/10.1016/j.eswa.2021.115540.

Purwarianti, A., & Yusliani, N. (2011). Sistem Question Answering Bahasa Indonesia untuk Pertanyaan Non-Factoid. *Jurnal Ilmu Komputer dan Informasi*, vol.4, no.1, 10-14. https://doi.org/10.21609/jiki.v4i1.151.

Santhi, B., & Brindha, G.R. (2019). Multinomial Naïve Bayes using Similarity Based Conditional Probability. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 36, issue 2, 1431-1441. https://doi.org/10.3233/JIFS-181009.

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *International Conference on Automation, Computational, and Technology Management (ICACTM)*, 593-596.

Sitepu, B. S., Munthe, I. R., & Harahap, Z. S. (2022). Implementation of Support Vector Machine Algorithm for Shopee Customer Sentiment Analysis. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol.7(2). https://doi.org/10.33395/sinkron.v7i2.11408.

Syahputra, R., Yanris, G. J., & Irmayani, D. (2022). SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol.7(2). https://doi.org/10.33395/sinkron.v7i2.11430.

Van-Tu, N., & Anh-Cuong, Le. (2016). Improving Question Classification by Feature Extraction and Selection. *Indian Journal of Science and Technology*, vol.9(17). doi : 10.17485/ijst/2016/v9i17/93160.

Zai, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2018). A Chi-Square Statistics Based Feature Selection Method in Text Classification. *IEEE 9th International Conference on Software Engineering and Service Science*.

Zulqarnain, M., Khalaf Zager Alsaedi, A., Ghazali, R., Ghouse, MG., Sharif, W., Aida Husaini, N. (2021). A comparative analysis on question classification task based on deep learning approaches. *PeerJ Computer Science* 7:e570 https://doi.org/10.7717/peerj-cs.570.

*name of corresponding author