INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Feature Selection Technique to Improve the Instances Classification Framework Performance for Quran Ontology

Yuli Purwati [a], Fandy Setyo Utomo [a,*], Nikmah Trinarsih [a], Hanif Hidayatulloh [a]

[a] *Department of Informatics Engineering, Faculty of Computer Science, Universitas AMIKOM Purwokerto, Purwokerto, 53127, Indonesia*
Corresponding author: [*]fandy_setyo_utomo@amikompurwokerto.ac.id

Abstract*—The Al-Quran is the sacred book of Muslims, and it provides God's word in the form of orders, instructions, and guidelines for people to follow to have happy lives both here and in the afterlife. Several earlier research has used ontologies to store the knowledge found in the Quran. The previous study focused on extracting the relationship between classes and instances or the "is-a relation" by classifying instances based on the referenced class. Based on the performance testing of the instances classification framework, the test results show that Support Vector Machine (SVM) with Term Frequency-Inverse Document Frequency (TF-IDF) and stemming operation had dropped the accuracy value to 65.41% when the test data size was increased to 30%. Likewise, with BPNN with TF-IDF and stemming operations. In the Indonesian Quran translation dataset with a test data size of 30%, the accuracy value drops to 57.86%. Instances classification based on the thematic topics of the Qur'an aims to connect verses (instances) to topics (classes) to get an overall picture of the topic and provide a better understanding to users. This study aims to apply the feature selection technique to the instances classification framework for the Al-Quran ontology and to analyze the impact of applying the feature selection technique to the framework with a small dataset and training data. The instances classification framework in this study consists of several stages: text-preprocessing, feature extraction, feature selection, and instances classification. We applied Chiq-Square as a technique to perform feature selection. SVM and BPNN as a classifier. Based on the experiment results, it can be concluded that the feature selection implementation using Chi-Square increases the value of precision, f-measure, and accuracy on the test data size from 40% to 60% in all datasets. The feature selection using Chi-Square and SVM classifier provides the highest precision value with a test data size of 60% on the Tafsir Quran dataset from the Ministry of Religious Affairs Indonesia: 64.36%. Furthermore, the feature selection implementation and BPNN classifier also increase the highest accuracy value with a test data size of 60% in the Quranic Tafsir dataset from the Ministry of Religion of the Republic of Indonesia: 63.09%.

***Keywords*— Ontology population; Chi-Square; machine learning; SVM; BPNN.**

Manuscript received 9 Sep. 2022; revised 21 Oct. 2022; accepted 11 Nov. 2022. Date of publication 30 Jun. 2023.

I. INTRODUCTION

The Al-Quran is the sacred book of Muslims, and it provides God's word in the form of orders, instructions, and guidelines for people to follow to have happy lives both here and in the afterlife. Several earlier research has used ontologies to store the knowledge found in the Quran. According to our literature review, two methods can be employed to generate ontologies, namely manual (non-automated processes) and automated processes. An ontology population is a technical term for this automated procedure [1]. Studying concepts, relationships, and examples from natural language texts and then adding them to an ontology is the process of ontology population [2]-[4].

The study by Utomo, Suryana, and Azmi [5] focused on extracting the relationship between classes and instances or the "*is-a relation*" by classifying instances based on the referenced class. Based on the performance testing of the instances classification framework, their test results show that Support Vector Machine (SVM) with Term Frequency-Inverse Document Frequency (TF-IDF) and stemming operation has the highest classification accuracy rate of up to 70.75% on the Indonesian Quran translation dataset with 20% of the test data size. However, when the test data size was increased to 30%, the accuracy value dropped to 65.41%. Likewise, with BPNN with TF-IDF and stemming operations. In the Indonesian Quran translation dataset with a test data size of 20%, the accuracy value reaches 61.32%. However, when the test data size increases to 30%, the accuracy value drops to 57.86%. Instances classification based on the

thematic topics of the Qur'an aims to connect verses (instances) to topics (classes) to get an overall picture of the topic and provide a better understanding to users [6],[7]. In addition, classifying the instance is to reduce the search space by identifying information relevant to a particular topic [8],[9]. The Quran ontology should store knowledge regarding thematic topics, verses, and their interpretations to help people gain better insight and understanding of the contents of the Quran.

Studies on the verses classification and the Qur'an interpretations based on thematic topics have been carried out by several previous researchers. This research classifies verses and interpretations into a single topic/class. The next part describes the results of recent studies in this research field, their findings, language and datasets, classifiers, research relations with ontologies, and measurement metrics.

A dataset was built with an English Quran translation and their Tafsir containing the surah Al-Baqarah and Al-An'am [10]-[13]. The Quranic verses in the two surahs are classified into three classes: faith, worship, and morality. They use k-NN, SVM, Naive Bayes, and decision trees (J48) as classifiers. However, their classification is not intended for the development of the ontology of the Quran. The study conducted by Adeleke et al. [13] used a Group-Based Feature Selection (GBFS) as one of the datasets with feature selection techniques: information gain (IG), chi-square (CH), Pearson correlation coefficient (PCC), Relief, and Correlation-based Feature selection (CFS) for the classification of verses of the Qur'an. The test results show that the highest classification accuracy is obtained from the SVM classifier by applying the CFS feature selection technique on the GBFS dataset: 94.5%, followed by applying IG and CH feature selection on the same classifier and dataset with an accuracy of 93.1%. In addition, the evaluation results show that the application of IG, CH, and CFS feature selection with the Naive Bayes, SVM, k-NN, and J48 classifiers on the English Quran translation, Tafsir, and GBFS datasets has a higher level of accuracy compared to the classification process without using feature selection techniques.

Furthermore, the study by Adeleke and Samsudin [12] used IG, CFS, and feature selection hybrid techniques with IG and CFS (IG-CFS) to classify the Qur'anic text. The test results show that applying IG, CFS, and IG-CFS can improve accuracy than classification without feature selection techniques. The implementation of CFS and IG-CFS with the SVM classifier obtained the highest accuracy value: 94.5% on the combined dataset between translation and Tafsir. In the subsequent research, in addition to the IG technique, the study conducted by Adeleke et al. [11] also used the CH technique in the classification process. The test results show that applying the two feature selection techniques can improve classification accuracy compared to the classification process without feature selection. The CH and IG implementation with the Naive Bayes classifier on the combined dataset between translation and interpretation has the highest accuracy: 93.9%. Similar to the study conducted by Adeleke and Samsudin [12], the feature selection technique used in research by Adeleke et al. [10] also uses several feature selection models: Chi-Square CH), CFS, and Hybrid CH-CFS for the Qur'anic text classification. The evaluation results are the same as other research findings that using feature selection techniques can improve accuracy compared to the classification process without feature selection techniques. Based on the test results on the accuracy measurement, the application of SVM with CFS and CH-CFS in the translation and interpretation dataset group has the highest accuracy: 93.6%.

The feature selection techniques implementation for verses classification and Quran interpretations based on specific thematic topics can improve the accuracy of the classification results [10]-[13]. Unlike the previous research that has been described, the study conducted by Utomo, Suryana, and Azmi [5] focuses on analyzing the impact of the stemmer algorithm on the instances classification framework for the development of the Indonesian Quran ontology. The instances are the Quran verses. They used the Indonesian Quran translation, Quraish Shihab's Tafsir, and the Indonesian Ministry of Religion's Tafsir Quran as datasets. Some of the Qur'an surahs used as datasets: Al-Baqarah, Ali Imran, An-Nisa', Al-An'am, Al-A'raf, At-Taubah, An-Nahl, and Taha. Classifiers used in their study: k-NN, SVM, and Backpropagation Neural Network (BPNN). They have not implemented the feature selection technique into the instances classification framework. The evaluation results of the instances classification framework using accuracy measurement metrics have been described in the previous paragraph. Based on the results of their evaluation, this study aims to apply the feature selection technique to the instances classification framework developed by Utomo, Suryana, Azmi [5] for the Al-Quran ontology and to analyze the impact of applying the feature selection technique to the framework with a small dataset and training data. The main contributions of our research are (1) developing an instances classification framework by applying feature selection techniques to increase classification performance; and (2) providing knowledge on the impact of these techniques on the framework's performance.

II. Materials and Method

This section presents our proposed framework, datasets, experimental settings, and test scenarios used in our research.

A. The Proposed Framework

The instances classification framework in the study conducted by Utomo, Suryana, and Azmi [5] consists of several stages: text-preprocessing, feature extraction, and instances classification. In this study, we propose implementing feature selection techniques into the framework adopted by Utomo, Suryana, and Azmi [5]. Fig. 1 describes the instances classification framework that we propose in this study.

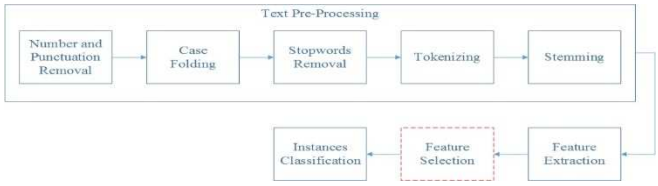

Fig. 1 The proposed instances classification framework

According to Fig. 1, automatic text pre-processing consists of several stages: number and punctuation Removal, case folding, stopwords removal, tokenizing, and stemming. In the

stop-word removal phase that follows the case-folding technique, common words thought to lack significance are eliminated from the sentences. This phase utilized Tala's 757-word stop-word list [14]. After the tokenizing phase, a stemming operation is carried out. Our study adopts the Sastrawi stemmer algorithm used in research by Utomo, Suryana, and Azmi [5]. Stemming is removing affixes from a word to convert it to its base form. The Sastrawi stemmer results from Nazief and Adriani's algorithm's [15] optimization and Confix Stripping (CS) algorithm, Enhanced Confix Stripping (ECS) algorithm, and Modified ECS algorithm by two previous studies (see [16], [17]) improved this stemmer. Sastrawi has implemented the solution from [18] to handle the suffix removal failure to improve stemming outcomes. This method addressed the suffix removal issue caused by Nazief and Adriani's stemmer.

Next, the text enters the feature extraction phase after the stemming phase using the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) model. BoW is an extraction model representing text as an unstructured word collection that disregards grammatical structure [19]. The feature selection operation is performed on the sparse matrix data representing the BoW model. The output of this operation is the selected BoW features that will convert into the TF-IDF model. The TF-IDF is a statistical method that represents the word meaning in a corpus by comparing a word's occurrence in one text to its occurrence in another document [20].

Feature selection is identifying the optimal subset of the original set of features. The selected subset's characteristics should be informative and discriminating. In other words, feature selection aims to reduce the number of features utilized to describe a dataset. Feature selection aids in reducing storage requirements, avoiding the over-fitting issue, and maintaining performance precision by minimizing excessive noise [21]–[26]. In this study, we use Chi-Square as a feature selection technique. The Chi-Square method is a hypothesis evaluation technique utilized to determine whether two variables are independent. A higher Chi-square value indicates a more significant variance between the two variables, and a lower correlation between them implies a stronger degree of independence [27]. Equation 1 shows the Chi-Square test formula [27]–[30].

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

where N = Total number of documents in the corpus; A = Number of documents in class c_i that contain the term t_k; B = Number of documents that contain the term t_k in other classes; C = Number of documents in class c_i that do not contain the term t_k; D = Number of documents that do not contain the term t_k in other classes.

Furthermore, we use the Support Vector Machine (SVM) and Backpropagation Neural Network (BPNN) algorithms in the instances classification stage. In this study, we classify Quranic verses into Morals, Al-Quran, and Previous Nation. Each verse is classified into a single class. The materials and methods section describes the datasets, experimental settings, and test scenarios used in this study.

B. Dataset

We used multiple data sources to build the research dataset. Table I describes the data, sources, and functions used in this study. Indonesian Quran translation, Quraish Shihab Tafsir, and Tafsir Quran from The Ministry of Religious Affairs Indonesia were used in this study as training data and test data in the experimental phase. Furthermore, thematic topics from the Quran Cordoba are used as the target class for classifying instances. This study uses three classes: Morals, Al-Quran, and Previous Nations. Meanwhile, the root word dictionary from Kateglo is used in the stemming phase for checking root words. Finally, the Indonesian stop word list is used in the text pre-processing phase to eliminate words that have no meaning in the sentence.

TABLE I
DATASET COLLECTION

No.	Data	Source
1.	Indonesian Quran Translation	International Quranic project: Tanzil - Quran Navigator (http://tanzil.net)
2.	Quraish Shihab Tafsir	
3.	Quran Tafsir	The Ministry of Religious Affairs Indonesia (https://quran.kemenag.go.id/)
4.	Thematic Topics	The Quran Cordoba thematic index
5.	Root word dictionary	Kateglo (https://kateglo.com/)
6.	Indonesian Stop word list	Indonesian stop word list from Tala [14]

After the data on Indonesian Quran translation and Tafsir Quran were obtained, then the data was used in this study to build a corpus based on thematic topics. The number of Quranic verses related to the thematic topics is presented in Table II.

TABLE II
THEMATIC TOPICS AND QURAN VERSES AMOUNT

Topic ID	Topic Name	Quran Verses Amount
1	Morals	218
2	Al-Quran	183
3	Previous Nations	127
	Total	**528**

Based on Table II, this study uses 528 Indonesian Quran translations, 528 interpretations of the Qur'anic verses from Quraish Shihab, and 528 interpretations of the Qur'anic verses from the Ministry of Religious Affairs Indonesia. In Table II, several chapters of the Qur'an are used as a dataset: Al-Baqarah, Ali Imran, An-Nisa', Al-An'am, Al-A'raf, At-Taubah, An-Nahl, Taha. Next, Table III describes the surahs of the Qur'an, the number of Qur'an verses in the surah, and the thematic topics used to develop the dataset.

TABLE III
SURAH'S NAME, NUMBER OF VERSES, AND THEMES

	Morals	Al-Quran	Prev. Nations	+
Al-Baqarah	51	59	13	**123**
Ali-Imran	40	29	28	**97**
An-Nisa'	47	25	12	**84**
Al-An'am	13	20	10	**43**
Al-A'raf	21	16	37	**74**
At-Taubah	28	8	4	**40**
An-Nahl	14	18	5	**37**
Taha	4	8	18	**30**

Based on Table III, the "+" column provides information on the number of verses in the surah used to build the corpus based on the three thematic topics. We adopted Al-Quran Cordoba's categorization of Quran verses into a single thematic area for this study.

C. Experimental Setup

We are developing an operational framework for the instances classification and experiments in the Python programming environment. Fig. 2 describes the operational framework that we used in this study. The operational framework we propose is a form of development of the operational framework used by Utomo, Suryana, and Azmi [5]. The difference between our operational framework and the operational framework applied by Utomo, Suryana, and Azmi [5] is that we applied the feature selection technique after the feature extraction phase in this study.

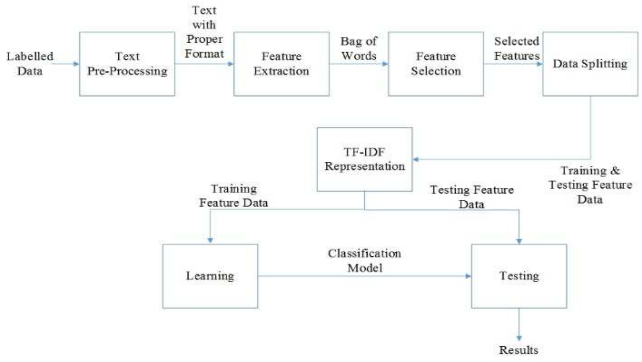

Fig. 2 The operational framework

According to Fig. 2, the inputs are Indonesian Quran translation, Quraish Shihab's Tafsir, and Quranic Tafsir from the Ministry of Religious Affairs Indonesia. All data from the three sources have been labeled as one of the target classes. Next, the data enters the text pre-processing stage. The phase sequence in this stage is described in Fig. 1. The purpose of the text pre-processing stage is to transform the text into an appropriate format.

Furthermore, feature extraction is carried out into the sparse matrix of features (BoW representation), and feature selection is performed using the Chi-Square technique on the BoW sparse matrix. After feature selection, the selected features are divided into training and test data features. Then the data are transformed into the TF-IDF model. Finally, the classification process is performed by BPNN and SVM algorithms against training and test data.

D. Test Scenario

We applied several test data sizes to investigate and analyze the impact of feature selection techniques on the instances classification performance with different classifiers: BPNN and SVM. The size of the test data for each thematic topic is shown in Table IV.

TABLE IV
THE TEST DATA SIZE FOR EACH THEMATIC THEMES

Test Data Size	Morals (Data)	Al-Quran (Data)	Prev. Nations (Data)	Sum (Data)
40%	88	73	51	212
50%	109	91	64	264
60%	131	110	76	317

Three test scenarios are presented in Table IV to study and evaluate the impact of feature selection approaches on the instances classification performance. We applied the precision, f-measure, and accuracy metric to evaluate the classification outcomes in this study.

III. Results and Discussion

We began our experiment by utilizing the Indonesian Quran translation (IQT) corpus. The test data size outlined in Table IV was utilized in this experiment. The comparison of instances classification performance between the framework used in research by Utomo, Suryana, and Azmi [5] and the proposed framework in our study using the precision metric is presented in Fig. 3. In contrast, Fig. 4 describes the measurement results using the F-Measure Metric, and Fig. 5 represents the measurement results using the accuracy metric.

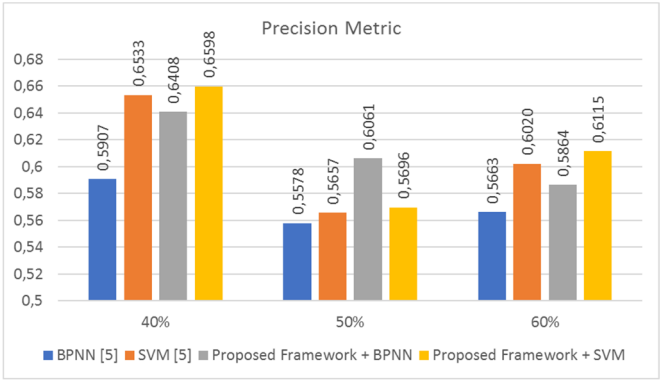

Fig. 3 Precision metric: Indonesian Quran Translation Corpus

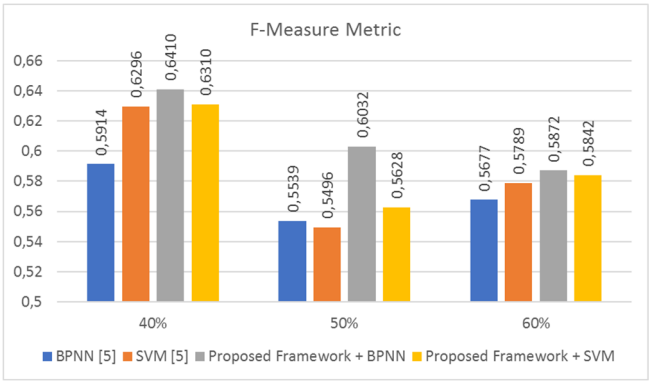

Fig. 4 F-measur e metric: Indonesian Quran Translation Corpus

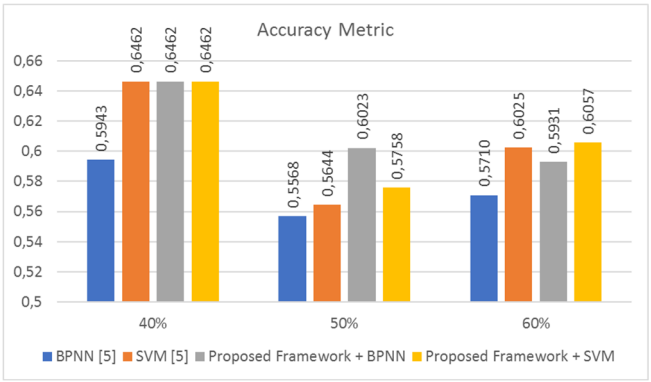

Fig. 5 Accuracy Metric: Indonesian Quran Translation Corpus

According to Fig. 3 to Fig. 5, we can conclude that the feature selection technique implementation could improve the instances classification performance based on evaluation

using precision, f-measure, and metric accuracy measurements on all test data sizes and both classifiers: BPNN and SVM. Feature selection in the instances classification framework with SVM and test data size of 60% has the highest precision value: 0.6115. Likewise, measurements using the Accuracy metric achieved the highest result of 0.6057 under similar conditions.

Furthermore, Fig. 6 to 8 explain the results of evaluating the feature selection technique implementation in the instances classification process using Quraish Shihab's Tafsir data.

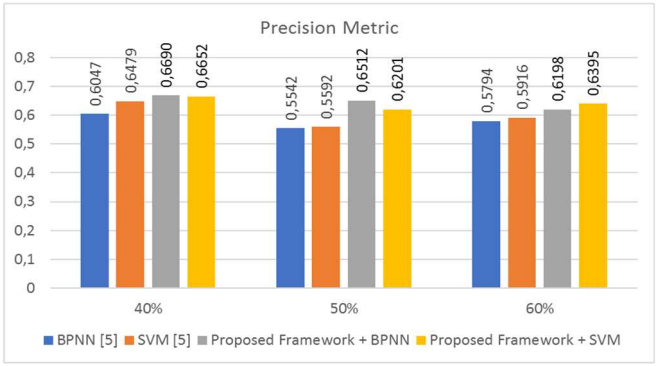

Fig. 6 Precision metric: Quraish Shihab's Tafsir Corpus

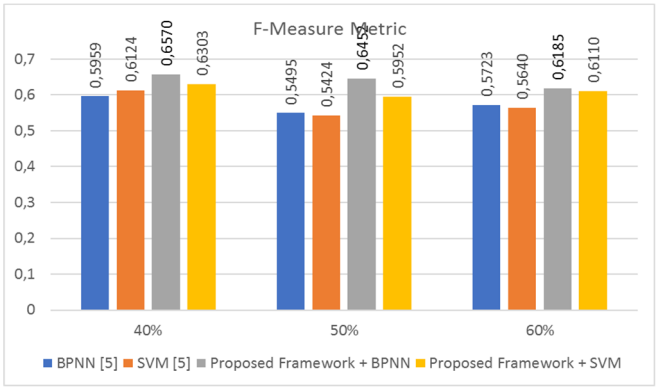

Fig. 7 F-measure metric: Quraish Shihab's Tafsir Corpus

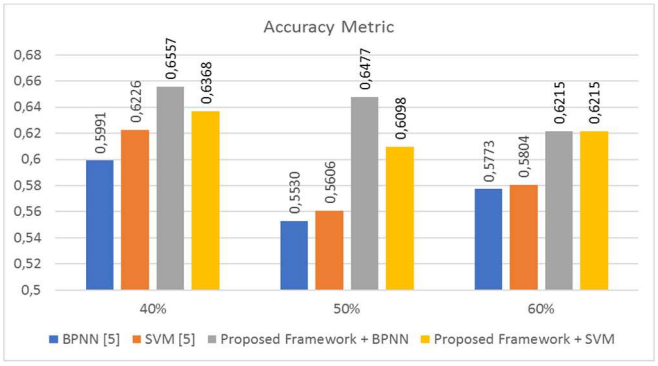

Fig. 8 Accuracy Metric: Quraish Shihab's Tafsir Corpus

According to Fig. 6 to Fig. 8, we can conclude that the feature selection technique implementation in Quraish Shihab's Tafsir could improve the instances classification performance based on evaluation using precision, f-measure, and metric accuracy measurements on all test data sizes and both classifiers: BPNN and SVM. Feature selection in the instances classification framework with SVM and test data size of 60% has the highest precision value: 0.6395. Furthermore, at the test data size of 50%, the feature selection implementation on the instances classification framework with the BPNN classifier achieved the highest accuracy: 0.6477. Meanwhile, in the 60% test data size, the feature selection implementation with the BPNN and SVM classifiers achieved the best accuracy with 0.6215.

Furthermore, Fig. 9 to 11 present the results of evaluating the feature selection technique implementation in the instances classification process using the Tafsir Quran from the Ministry of Religious Affairs Indonesia.

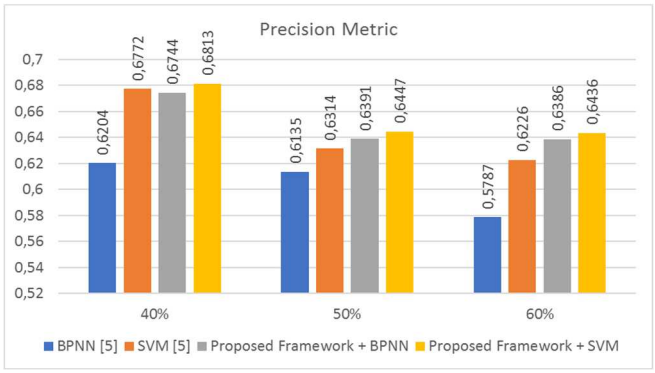

Fig. 9 Precision metric: the Ministry of Religious Affairs Indonesia Tafsir

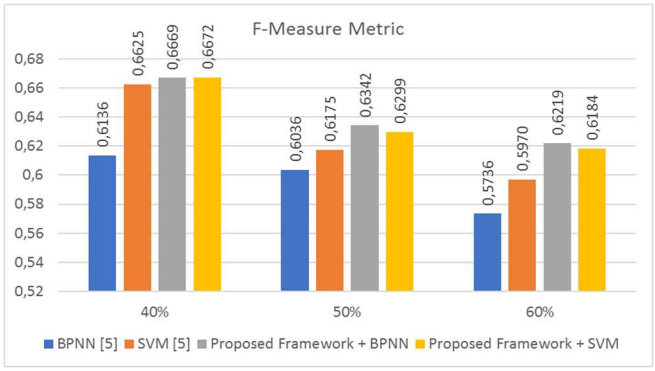

Fig. 10 F-measure metric: the Ministry of Religious Affairs Indonesia Tafsir

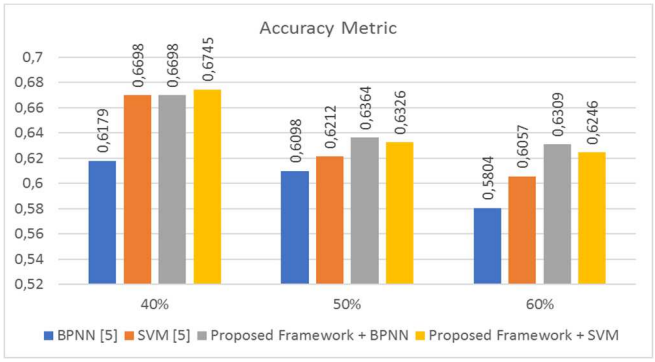

Fig. 11 Accuracy metric: the Ministry of Religious Affairs Indonesia Tafsir

According to Fig. 9 to Fig. 11, we can conclude that the feature selection technique implementation in Quran Tafsir from the Ministry of Religious Affairs Indonesia could improve the instances classification performance based on evaluation using precision, f-measure, and metric accuracy measurements on all test data sizes and both classifiers: BPNN and SVM. The feature selection technique implementation in the instances classification framework with the SVM classifier shows the highest evaluation results on precision metric measurements in all test data sizes based on Fig. 9. At 40%; precision reaches 0.6813. While on the test

data size of 50%, the precision reached 0.6447. Then at 60% test data size, the precision reaches 0.6436.

Furthermore, the feature selection technique implementation in the instances classification process using the BPNN classifier at 50% and 60% test data sizes also achieved the highest evaluation results based on the accuracy metric, as shown in Fig. 11. At 50% test data size, the accuracy reached 0.6364. While the test data size is 60%, the accuracy is 0.6309.

IV. Conclusion

Based on the study's results, it can be concluded that the feature selection implementation using Chi-Square increases the value of precision, f-measure, and accuracy on the test data size from 40% to 60% in all datasets. The feature selection using Chi-Square provides the highest precision value with a test data size of 60% on the Tafsir Quran dataset from the Ministry of Religious Affairs Indonesia: (1) the Proposed framework with BPNN = 0.6386; and (2) the proposed framework with SVM = 0.6436. Furthermore, the feature selection implementation also provides an increase in the highest accuracy value with a test data size of 60% in the Quranic Tafsir dataset from the Ministry of Religion of the Republic of Indonesia: (1) the proposed framework with BPNN = 0.6309; and (2) the proposed framework with SVM = 0.6246.

Acknowledgment

The authors thank Universitas AMIKOM Purwokerto for granting financial support to fund our research in 2022 with decree number 47/AMIKOMPWT/LPPM/15/IV/ 2022. Furthermore, we also would like to thank the AMIKOM Purwokerto Intelligent System (AI-SYS) research group for their assistance in this research.

References

[1] N. Suryana, F. S. Utomo, and M. S. Azmi, "Quran Ontology: Review on Recent Development and Open Research Issues," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 3, pp. 568–581, 2018.

[2] P. Cimiano, *Ontology Learning and Population from Text*, 1st ed. New York, NY: Springer US, 2006.

[3] R. Witte, R. Krestel, T. Kappler, and P. C. Lockemann, "Converting a Historical Architecture Encyclopedia into a Semantic Knowledge Base," *IEEE Intelligent Systems*, vol. 25, no. 1, pp. 58–67, 2010.

[4] J. A. Reyes and A. Montes, "Learning Discourse Relations from News Reports: An Event-driven Approach," *IEEE Latin America Transactions*, vol. 14, no. 1, pp. 356–363, 2016.

[5] F. S. Utomo, N. Suryana, and M. S. Azmi, "Stemming impact analysis on Indonesian Quran translation and their exegesis classification for ontology instances," *IIUM Engineering Journal*, vol. 21, no. 1, pp. 33–50, 2020.

[6] R. Ismail, Z. Abu Bakar, and N. Abd Rahman, "Extracting knowledge from english translated quran using NLP pattern," *Jurnal Teknologi*, vol. 77, no. 19, pp. 67–73, 2015.

[7] N. K. Farooqui and M. F. Noordin, "Knowledge exploration: Selected works on Quran ontology development," *Journal of Theoretical and Applied Information Technology*, vol. 72, no. 3, pp. 385–393, 2015.

[8] S. K. Hamed and M. J. Ab Aziz, "A question answering system on Holy Quran translation based on question expansion technique and Neural Network classification," *Journal of Computer Science*, vol. 12, no. 3, pp. 169–177, 2016.

[9] B. Baharudin, L. H. Lee, K. Khan, and A. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification," *Journal of Advances in Information Technology*, vol. 1, no. 1, pp. 4–20, 2010.

[10] A. Adeleke, N. A. Samsudin, Z. A. Othman, and S. K. Ahmad Khalid, "A two-step feature selection method for quranic text classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 730–736, 2019.

[11] A. Adeleke, N. Samsudin, A. Mustapha, and S. Ahmad Khalid, "Automating quranic verses labeling using machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 925–931, 2019.

[12] A. Adeleke and N. Samsudin, "A Hybrid Feature Selection Technique for Classification of Group-based Holy Quran Verses," *International Journal of Engineering & Technology*, vol. 7, no. 4.31, pp. 228–233, 2018.

[13] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. M. Nawi, "A group-based feature selection approach to improve classification of Holy Quran verses," *Advances in Intelligent Systems and Computing*, vol. 700, pp. 282–297, 2018.

[14] F. Z. Tala, "A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia," Universiteit van Amsterdam, 2003.

[15] J. Asian, "Effective Techniques for Indonesian Text Retrieval," RMIT University, 2007.

[16] R. Kusumaningrum, S. Adhy, and S. Suryono, "WCLOUDVIZ: Word Cloud Visualization of Indonesian News Articles Classification based on Latent Dirichlet Allocation," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 16, no. 4, pp. 1752–1759, 2018.

[17] I. G. M. Darmawiguna, G. A. Pradnyana, and G. S. Santyadiputra, "The Development of Integrated Bali Tourism Information Portal using Web Scrapping and Clustering Methods," *Journal of Physics: Conference Series*, vol. 1165, no. 1, pp. 1–10, 2019.

[18] A. Z. Arifin, I. P. A. D. Mahendra, and H. T. Ciptaningtyas, "Enhanced Confix Stripping Stemmer and Ants Algorithm For Classifying News Document in Representation of Textual," in *The 5th International Conference on Information & Communication Technology and Systems*, pp. 149–158, 2009.

[19] M. J. Schneider and S. Gupta, "Forecasting sales of new and existing products using consumer reviews: A random projections approach," *International Journal of Forecasting*, vol. 32, no. 2, pp. 243–256, 2016.

[20] G. Chen and L. Xiao, "Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods," *Journal of Informetrics*, vol. 10, no. 1, pp. 212–223, 2016.

[21] P. Agrawal, H. F. Abutarboush, T. Ganesh, and A. W. Mohamed, "Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019)," *IEEE Access*, vol. 9, pp. 26766–26791, 2021.

[22] C. P. Vandana and A. A. Chikkamannur, "Feature selection: An empirical study," *International Journal of Engineering Trends and Technology*, vol. 69, no. 2, pp. 165–170, 2021.

[23] M. Qaraad, S. Amjad, I. I. M. Manhrawy, H. Fathi, B. A. Hassan, and P. El Kafrawy, "A Hybrid Feature Selection Optimization Model for High Dimension Data Classification," *IEEE Access*, vol. 9, pp. 42884–42895, 2021.

[24] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to Multi-Objective Feature Selection: A Systematic Literature Review," *IEEE Access*, vol. 8, pp. 125076–125096, 2020.

[25] H. Nematzadeh, R. Enayatifar, M. Mahmud, and E. Akbari, "Frequency based feature selection method using whale algorithm," *Genomics*, vol. 111, no. 6, pp. 1946–1955, 2019.

[26] L. Zhu, S. He, L. Wang, W. Zeng, and J. Yang, "Feature selection using an improved gravitational search algorithm," *IEEE Access*, vol. 7, pp. 114440–114448, 2019.

[27] H. Kang, G. Liu, Z. Wu, Y. Tian, and L. Zhang, "A Modified FlowDroid Based on Chi-Square Test of Permissions," *Entropy*, vol. 23, no. 2, p. 174, 2021.

[28] S. Bahassine, A. Madani, M. Al-sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, 2020.

[29] Y. D. Setiyaningrum, A. F. Herdajanti, C. Supriyanto, and Muljono, "Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm," in *International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 78–81, 2019.

[30] S. T. Ikram and A. K. Cherukuri, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University - Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.