Word embedding and imbalanced learning impact on Indonesian Quran ontology population

Fandy Setyo Utomo[1], Yuli Purwati[2], Mohd Sanusi Azmi[3], Lulu Shafira[2], Nikmah Trinarsih[2]
[1]Department of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia
[2]Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto, Purwokerto, Indonesia
[3]Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Article Info

Article history:

Received Oct 19, 2024
Revised Mar 10, 2025
Accepted Mar 26, 2025

Keywords:

Indonesian Quran interpretation
Machine learning
Ontology learning
Smote
Word2vec

ABSTRACT

This research addresses limitations in Quranic instance classification, exceptionally high dimensionality, lack of semantic relationships in the term frequency-inverse document frequency (TF-IDF) technique, and imbalanced data distribution, which reduce prediction accuracy for minority classes. This study investigates the impact of word embedding and imbalance learning techniques on instance classification frameworks using Indonesian Quran translation and Tafsir datasets to handle previous research limitations. Four classification frameworks were built and evaluated using accuracy and hamming loss metrics. The results show that the synthetic minority oversampling technique (SMOTE) technique, TF-IDF model, and logistic regression classifier provide the best accuracy results of 62.74% and a hamming loss score of 0.3726 on the Quraish Shihab Tafsir dataset. This is better than the performance of previous classifiers backpropagation neural network (BPNN) and support vector machine (SVM) used in the previous framework, with accuracies of 59.91% and 62.26%, respectively. Logistic regression can also provide the best classification results with an accuracy of 67.92% and a hamming loss of 0.3208 using the previous framework. These results are better than the performance of the previous classifiers BPNN and SVM used in the previous framework, with accuracies of 62.26% and 66.98%, respectively. TF-IDF feature extraction outperforms word2vec in instance classification results due to its superior support under limited dataset conditions.

Corresponding Author:

Yuli Purwati
Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto
53127 Purwokerto, Banyumas, Indonesia
Email: yulipurwati@amikompurwokerto.ac.id

1. INTRODUCTION

The Qur'an, as a holy book, contains knowledge and scientific facts. Previous research has encoded the information found in the Quran into an ontology structure. Our literature research indicates that ontology development can be achieved through two methods: manual (non-automated) and automatic. This automated technique is referred to as the ontology population [1]. Ontology population refers to acquiring knowledge about concepts, relations, and instances from text written in natural language and subsequently incorporating this knowledge into an ontology [2]-[5]. The Quran ontology population approach can be performed using several methods: rule-based, natural language processing (NLP), statistics/machine learning, and hybrid approaches [1].

Several previous researchers have conducted studies regarding the instances classification for the Qur'an based on their thematic topics. The research conducted by [6]-[9] focuses on classifying instances with the English Quran translation dataset. The techniques employed for feature extraction include bag of words (BoW) and term frequency-inverse document frequency (TF-IDF). The investigation carried out by [9] employs random under-sampling (RUS) and synthetic minority oversampling technique (SMOTE) to address data imbalance. Furthermore, the research conducted by [10]-[13] focuses on classifying instances in the Arabic Quran dataset. The study [10] uses the diffusion frechet function for the feature extraction technique, while other studies use BoW and TF-IDF. To handle data imbalance in the study [13], they use SMOTE. Several previous researchers have researched the instances classification in the Indonesian Quran translation dataset. Studies conducted by [14]-[18] used the TF-IDF technique for feature extraction on the Indonesian Quran translation with single-label classification.

The work performed by [17] aims to extract "is-a relations" between classes and instances by classifying instances according to the referenced class. The evaluation of the instance classification framework's performance reveals that the support vector machine (SVM), employing TF-IDF and stemming methods, attains the highest classification accuracy of 70.75% on the Indonesian Quran translation dataset with a test data ratio of 20%. The subsequent study by [18] focused on enhancing the performance of the classification framework utilized by [17] by applying the chi-square feature selection method to minimize the dimension size. The test results indicate that feature selection methods can improve the precision and accuracy of instance classification outcomes when the test data size is configured at 60%. The SVM classifier attained a precision of 64.36% when utilized on the Indonesian Tafsir Al-Quran dataset sourced from the Ministry of Religion of the Republic of Indonesia. Furthermore, the backpropagation neural network (BPNN) classifier attained an accuracy of 63.09%.

The two previous studies in [17], [18] employed the TF-IDF term weighting technique within an instance classification framework. The TF-IDF term weighting technique has several limitations. Firstly, it is vulnerable to the issue of large dimensions in the text domain, adversely affecting classifiers' performance [19]. Additionally, it does not consider information about meaning, membership relationships, or semantics between words/documents in the data source [20]-[22]. Further, this method proves to be inefficient for the task of text classification when dealing with imbalanced data distribution [23]. Furthermore, there was an imbalanced data distribution in the research conducted by [17], [18]. Both studies provide three classification output targets: morals, Al-Quran, and previous nations. Nevertheless, the distribution of data utilized as a dataset in the three classes is imbalanced. The morals class has 218 data, the Al-Quran class has 183 data, and the previous nations class has 127 data. This will lead to significant differences in training and test data distribution in each class. Unfortunately, machine-learning algorithms often show inadequate results when there is a substantial imbalance in class occurrences. The presence of class imbalance challenges supervised models to accurately capture the distribution properties of skewed data, leading to reduced prediction accuracy for the minority classes [24]-[28].

Word embedding effectively overcomes the constraints associated with TF-IDF. Word embedding is a method utilized in NLP and machine learning that encodes words as dense vectors within a continuous vector space, with each word assigned to a high-dimensional vector [29]. This form of representation captures the semantic and syntactic information of words based on the context of their use from large textual content [30], [31]. Furthermore, the imbalance learning technique can be employed to address the issue of class data imbalance [24], [32], [33]. There are three approaches in imbalance learning that can be used to solve imbalance data problems: data-level, algorithm-level, and hybrid approaches [34], [35]. This research focuses on the data-level approach as a solution to data imbalance. This study employs and examines word embedding and imbalanced learning to mitigate the shortcomings of TF-IDF and data imbalance in instance classification systems, notably utilizing the Indonesian Quran translation and Tafsir dataset for the ontology population. The innovation resides in applying and examining these strategies to address both challenges.

2. METHOD

This section is structured as follows: section 2.1 discusses adopting the framework used in this study. Section 2.2 explains the dataset used in this research. Finally, the test scenario is defined in section 2.3.

2.1. Framework adopted

In this study, we adopt the instance classification framework from [18] to label each verse of the Indonesian Quran translation and its interpretation into a single thematic topic. The framework [18] has multiple stages: (1) text preprocessing: number and punctuation removal, case folding, stopword removal, and tokenization; (2) morphological analysis: stemming operation; (3) feature extraction; (4) feature selection; and (5) instance classification. The morphological analysis technique (stemming) used in this study is the same as the technique applied by [18] using the Sastrawi stemmer. Other researchers have also implemented this stemmer, such as the study conducted by [36]-[38] for text processing. Furthermore, this study's feature selection technique uses chi-square, as implemented by [18] in his research.

In this study, we use word2vec to implement the word embedding technique in the feature extraction stage. Word2vec is a neural network-driven word embedding method that depicts words in a continuous vector space, capturing both semantic and syntactic associations. This renders it very efficient for NLP tasks and text feature extraction across many applications [39]-[41]. Conventional feature selection techniques, like chi-square, information gain, and document frequency, consider solely the frequency of feature occurrences while neglecting feature semantics and part-of-speech attributes. Word2vec models generate vector representations of words that capture semantic meanings and are beneficial for numerous NLP tasks [42], [43]. Based on the limitations of traditional feature selection techniques, the instance classification framework adopted in this study must be modified. Section 2.3 will explain the modification. In the study [18], they used SVM and BPNN classifiers for text classification. However, in this study, we use SVM and BPNN, and also add logistic regression and random forest as classifiers.

2.2. Dataset collection

The research dataset was compiled from various sources, summarized in Table 1. The study used Quraish Shihab's, Indonesian Quran translation, and the Ministry of Religious Affairs' Tafsir Quran for training and testing. The classification target is the Quran Cordoba's thematic subjects, divided into the Quran, Historical Nations, and Morality. The Kateglo root word dictionary is simultaneously employed during the stemming procedure to verify root words. The text pre-processing phase employs the Indonesian stop word list to eliminate terms that are irrelevant to the text. Upon acquiring the data on Indonesian Quran translations and Tafsir, this study utilized the information to construct a corpus based on thematic subjects. Table 2 delineates the quantity of Quranic verses pertinent to the thematic subjects.

Table 1. Dataset collection

No.	Data	Source
1	Indonesian Quran translation	International Quranic project: Tanzil (http://tanzil.net)
2	Quraish Shihab Tafsir	International Quranic project: Tanzil (http://tanzil.net)
3	Quran Tafsir	The Ministry of Religious Affairs Indonesia (https://quran.kemenag.go.id/)
4	Thematic subjects	The Quran Cordoba thematic index
5	Root word dictionary	Kateglo (https://kateglo.com/)
6	Indonesian stop word list	Indonesian stop word list from Tala [44]

Table 2. Thematic subjects and number of Quranic verses

ID.	Thematic subject name	Number of Quranic verses
1	Morals	218
2	Al-Quran	183
3	Previous nations	127
	Sum total	528

Table 2 illustrates that this study employs 528 Qur'anic verses sourced from the Indonesian Ministry of Religious Affairs, alongside 528 Qur'anic verses from Quraish Shihab, and 528 translations of the Qur'an in Indonesian. The chapters of the Qur'an, including Al-Baqarah, Ali Imran, An-Nisa', Al-An'am, Al-A'raf, At-Taubah, An-Nahl, and Taha, are featured as part of the dataset presented in Table 2. Consequently, Table 3 delineates the thematic subjects that were employed to generate the dataset, the number of verses within each surah, and the surahs of the Qur'an. This research utilized the categorization of Quranic verses by Al-Quran Cordoba, organizing them into a cohesive thematic framework, as illustrated in Table 3.

Table 3. Surah, number of verses, and subjects

Surah name	Morals	Al-Quran	Prev. nations	Total
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

2.3. Test scenario

The training data size and test data size were 60% and 40%, respectively, in this investigation. The data size of each thematic topic can be elucidated as follows, based on the 40% test data size: (1) 88 data for the moral topic; (2) 73 data for the Al-Quran topic; and (3) 51 data for the topic of the previous nation, so that the total training data size is 212 data. Next, the SMOTE technique is implemented in our study to handle data imbalance. We use a model trained on the Indonesian Wikipedia corpus as a source model for implementing word embedding using the word2vec technique. This model can be downloaded via the Github website as follows: https://github.com/deryrahman/word2vec-bahasa-indonesia/tree/master. Each word will be represented as a 300-dimensional vector.

As described in section 2.1, the limitations of traditional feature selection techniques to handle feature extraction results using word2vec require modifications to the adopted instance classification framework. This research investigates the impact of word embedding and imbalanced learning on instances classification frameworks with the Indonesian Quran translation dataset for the ontology population. Based on these conditions, there are 4 test scenarios which can be explained as follows:

- Test scenario 1
  In test scenario 1, we build a framework for classifying instances from research [18]. We name it framework 1. The text pre-processing, morphological analysis, feature extraction, and feature selection stage use the same techniques as the research [18].
- Test scenario 2
  In the second test scenario, we modified the instance classification framework from the study [18]. The modification was made due to the limitations of traditional feature selection techniques in handling the results of word2vec feature extraction. The results of the framework modification are named framework 2 which consists of several phases: (i) text preprocessing: number and punctuation removal, case folding, stopword removal, and tokenization; (ii) morphological analysis: stemming operation; (iii) feature extraction; and (iv) instance classification. The feature extraction stage uses the word2vec technique.
- Test scenario 3
  In the third test scenario, we added an imbalance learning component to the instance classification framework [18] to handle the data imbalance for each dataset’s target class, so the framework has several work phases: (i) text preprocessing: number and punctuation removal, case folding, stopword removal, and tokenization; (ii) morphological analysis: stemming operation; (iii) imbalance learning; (iv) feature extraction; (v) feature selection; and (vi) instance classification. This work phase is called framework 3. We apply the SMOTE technique to the imbalance learning component and the TF-IDF technique for feature extraction.
- Test scenario 4
  Finally, in this fourth test scenario, we modified the instance classification framework [18] by adding an imbalance learning component using the SMOTE technique and performing feature extraction using word2vec. We did not use the traditional feature selection technique because it is limited in handling the results of word2vec feature extraction. This modification results, which we named framework 4 which consists of several work stages: (i) text preprocessing: number and punctuation removal, case folding, stopword removal, and tokenization; (ii) morphological analysis: stemming operation; (iii) feature extraction; (iv) imbalance learning; and (v) instance classification.

This study uses accuracy and hamming loss metrics to assess the efficacy of the instance classification system. The accuracy statistic is the proportion of real outcomes (including true positives and true negatives) relative to the total number of instances analyzed. It is a fundamental metric for evaluating a classifier’s performance, clearly indicating the model’s overall effectiveness [45]. This metric is particularly beneficial in balanced datasets, where the classes are represented equally. However, it may be misleading when the class distribution is imbalanced [46], [47]. Furthermore, the hamming loss quantifies the proportion of incorrect labels predicted by a classifier compared to the actual labels. Specifically, hamming loss is defined as the average number of misclassified labels per instance [48], [49]. A lower hamming loss indicates improved performance, as it indicates a reduction in the number of misclassifications.

3. RESULTS AND DISCUSSION

This section is organized as follows: section 3.1 examines the findings of employing word embedding and imbalance learning inside the instance classification framework of the IQT dataset. Subsequently, section 3.2 analyzes the findings from the Quraish Shihab dataset, while section 3.3 evaluates the conclusions of the Quran Tafsir dataset provided by the Ministry of Religious Affairs of Indonesia.

3.1. Investigation results on the IQT dataset

Figure 1 illustrates the evaluation results of the overall framework based on the established test scenarios, utilizing accuracy measures. Referring to Figure 1, it can be concluded that the instance classification framework from research [18] with the SVM classifier has the best accuracy compared to other frameworks, with 64.62% accuracy. This result is identical to their [18] previous research using a similar framework and classifier, which also had the same accuracy result. Then, Figure 2 shows the evaluation results using hamming loss. Referring to Figure 2, it can be concluded that the instance classification framework from research [18] with the SVM classifier has the lowest hamming loss value of 0.3538 compared to other frameworks.

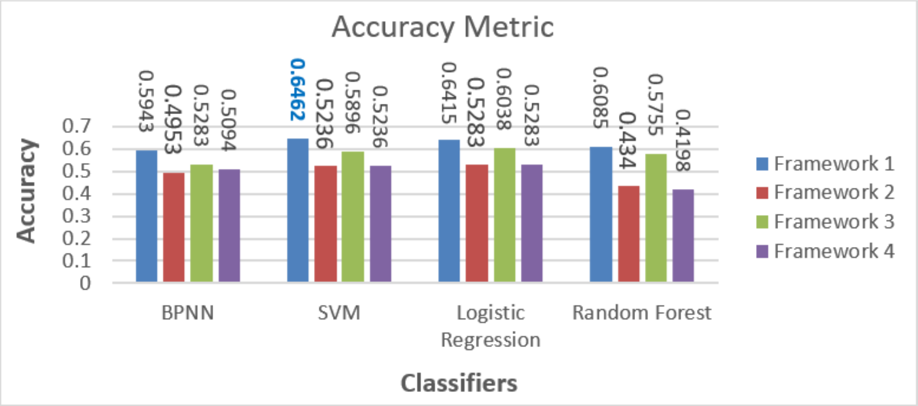


Figure 1. Framework performance evaluation with accuracy metric on IQT dataset

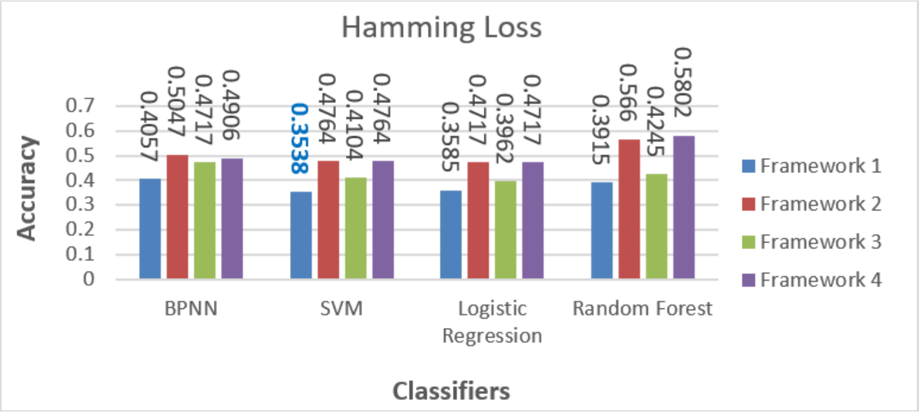
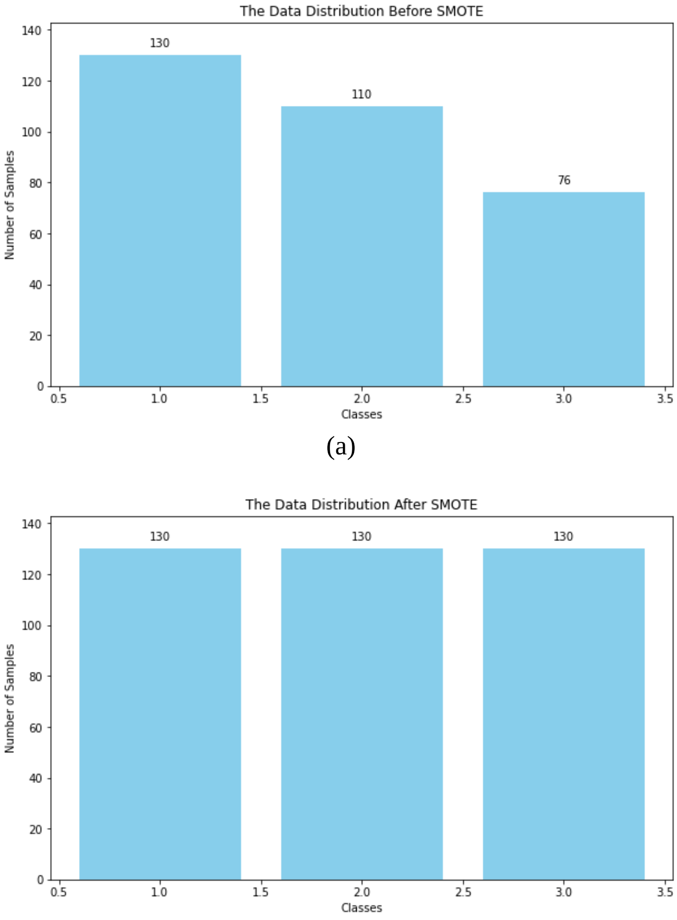


Figure 2. Framework performance evaluation with hamming loss metric on IQT dataset

Based on Figures 1 and 2, it can be explained that the impact of applying the SMOTE technique to handle data imbalance, together with TF-IDF for feature extraction in framework 3, has the best results compared to frameworks 2 and 4. It can be concluded that feature extraction utilizing TF-IDF has a more favorable impact on the framework than the word2vec technique, particularly regarding the accuracy and Hamming loss evaluation. Figure 3 shows a comparison between data before and after SMOTE is applied.

In Figure 3, class 1.0 denotes the moral class, class 2.0 represents the Al-Quran class, and class 3.0 pertains to the class of previous nations. Figure 3(a) illustrates the quantity of training data allocated to each target class before the execution of the SMOTE operation. The dataset is partitioned into 60% for training purposes and 40% designated for testing. The training data composition for each target class aligns with the total data presented in Table 2, as illustrated in Figure 3(a). The application of the SMOTE technique addresses the imbalance in training data across classes, ensuring that the minority class has an equal number of training instances as the majority class, totaling 130 instances, as illustrated in Figure 3(b).



(a)

(b)

Figure 3. Comparison of data amount: (a) before SMOTE; and (b) after SMOTE

3.2. Investigation results on the Quraish Shihab dataset

The Quraish Shihab dataset shows different evaluation results. Figure 4 describes the overall evaluation of the framework and classifier using the accuracy metric. According to Figure 4, it can be concluded that the instance classification framework 3 with the logistic regression classifier has the best accuracy compared to other frameworks, with 62.74% accuracy. This result is better than the instance classification framework 1 from research [18] using BPNN and SVM classifiers, which had accuracies of 59.91% and 62.26%, respectively. Then, Figure 5 shows the evaluation results using hamming loss.

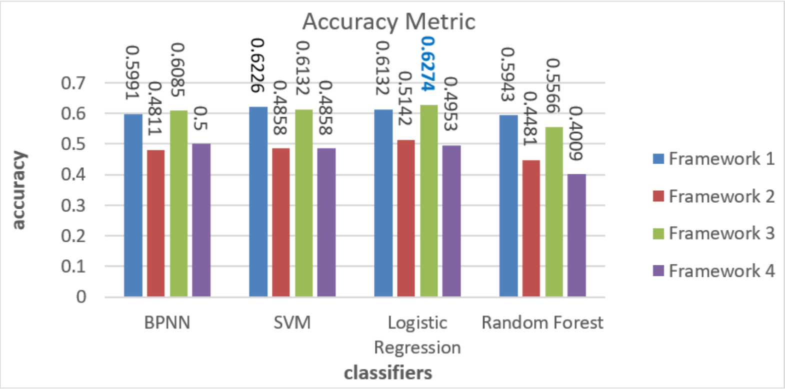


Figure 4. Framework performance evaluation with accuracy metric on Quraish Shihab dataset

Referring to Figure 5, it can be concluded that the instance classification framework 3 with the logistic regression classifier has the lowest hamming loss value of 0.3726 compared to other frameworks. Based on Figures 4 and 5, it can be explained that the impact of applying the SMOTE technique to handle data imbalance, together with TF-IDF for feature extraction in framework 3, has the best results compared to other frameworks. It can be concluded that feature extraction utilizing TF-IDF has a more favorable impact than the word2vec technique within the framework, particularly regarding the accuracy and hamming loss evaluation.

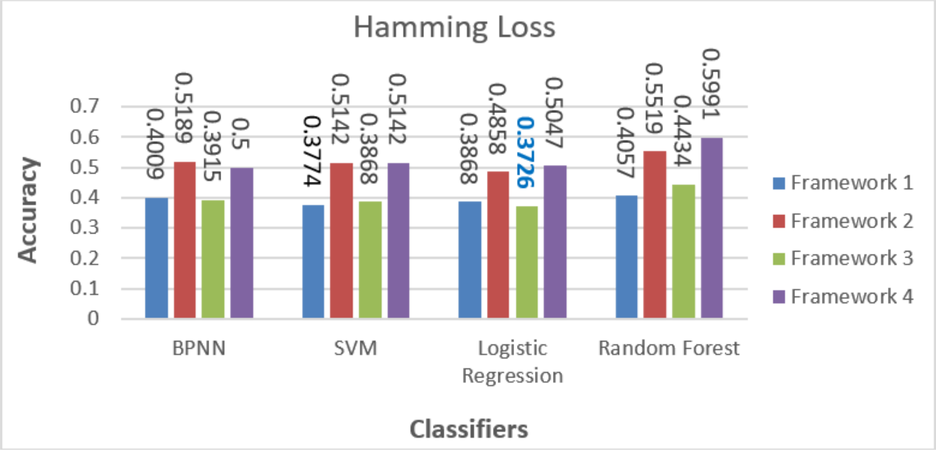


Figure 5. Framework performance evaluation with hamming loss metric on the Quraish Shihab dataset

3.3. Investigation results on the Quran Tafsir dataset from Ministry of Religious Affairs Indonesia

Figure 6 illustrates the evaluation outcomes of the comprehensive framework utilizing accuracy metrics derived from the established test scenarios. Referring to Figure 6, it can be concluded that the instance classification framework from research [18] with the logistic regression classifier has the best accuracy compared to other frameworks. This classifier performs better than the SVM and BPNN previously used by [18] in instances classification framework 1 with an accuracy of 62.26% and 66.98%, respectively.

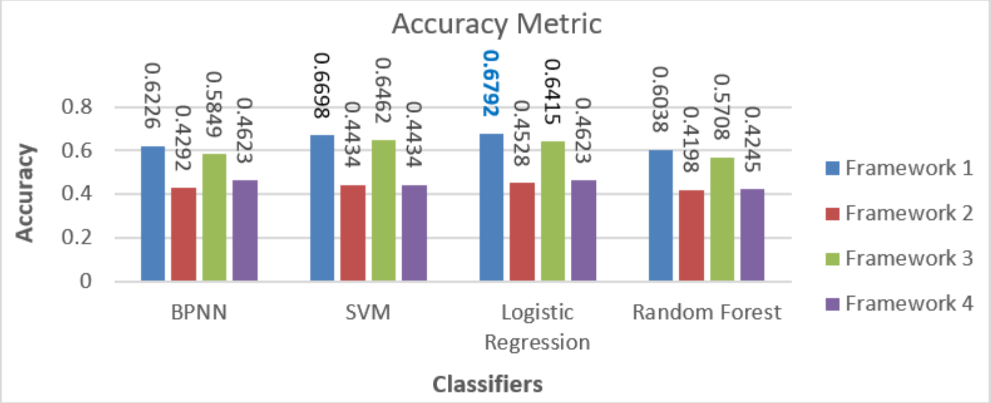


Figure 6. Framework performance evaluation with accuracy metric on the Quran Tafsir dataset

Moreover, Figure 7 shows the evaluation results using hamming loss. Based on Figures 6 and 7, the impact of applying the SMOTE technique to handle data imbalance, together with TF-IDF for feature extraction in framework 3, is the best compared to frameworks 2 and 4. The findings indicate that feature extraction through TF-IDF demonstrates a superior effect compared to the word2vec technique within the framework, particularly regarding accuracy and Hamming loss evaluation.

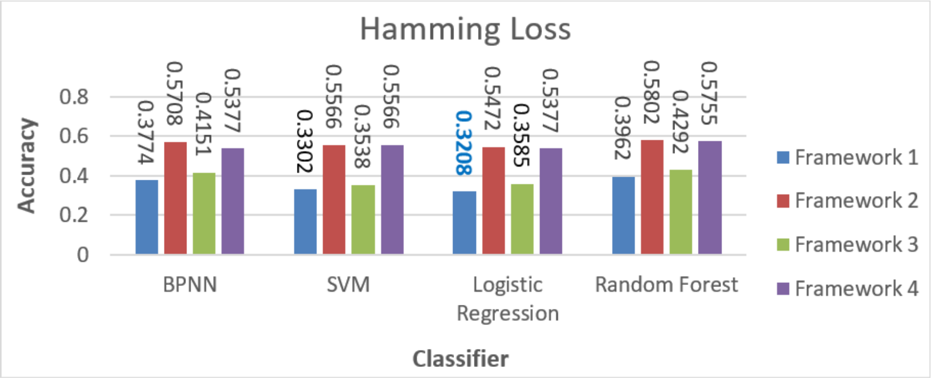


Figure 7. Framework performance evaluation with hamming loss metric on the Quran Tafsir dataset

4. CONCLUSION

The evaluation results of all frameworks and classifiers, assessed through accuracy and hamming loss metrics, indicate that the implementation of the SMOTE technique for imbalance learning in framework 1, combined with the TF-IDF model for feature extraction and the logistic regression classifier, achieved the highest accuracy of 62.74% and a hamming loss score of 0.3726 on the Quraish Shihab Tafsir dataset. This outcome surpasses the instance classification framework 1 from the previous study, utilizing BPNN and SVM classifiers, which achieved 59.91% and 62.26% accuracy. Logistic regression demonstrates the capability to yield optimal classification outcomes, achieving an accuracy of 67.92% and a hamming loss of 0.3208 in framework 1. The results surpass the performance of the earlier classifiers, BPNN and SVM, utilized in framework 1 from a previous study, which achieved accuracies of 62.26% and 66.98%, respectively. Additionally, it can be concluded that feature extraction through TF-IDF provides a more significant contribution compared to word2vec in enhancing the outcomes of instance classification, as evidenced by the performance of instance classification frameworks 1 and 3. In theory, word2vec is a word embedding technique that presents certain advantages over the TF-IDF feature extraction model. Our experiments indicate that the TF-IDF technique offers superior support compared to word2vec when classifying instances with limited dataset conditions.

ACKNOWLEDGEMENTS

The authors express their gratitude to Universitas Amikom Purwokerto for the financial support extended to our research in 2024, as outlined in decree number 046/AMIKOMPWT/P/11/VII/2024. We would also like to express our gratitude to the Amikom Purwokerto Intelligent System (AI-SYS) research group for their support in carrying out this investigation.

FUNDING INFORMATION

This research was funded by Universitas Amikom Purwokerto in 2024 with the decision letter number 046/AMIKOMPWT/P/11/VII/2024

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Fandy Setyo Utomo	✓	✓			✓	✓		✓	✓			✓		
Yuli Purwati			✓		✓	✓	✓	✓		✓				
Mohd Sanusi Azmi		✓		✓	✓	✓	✓							
Lulu Shafira						✓		✓		✓	✓		✓	✓
Nikmah Trinarsih						✓		✓		✓	✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

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: Algorithms, evaluation and applications*. 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. 58–66, Jan. 2010, doi: 10.1109/MIS.2010.17.

[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, Jan. 2016, doi: 10.1109/TLA.2016.7430101.

[5] J. A. Reyes-Ortiz, M. Bravo, and H. Pablo, “Web services ontology population through text classification,” in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016*, Oct. 2016, pp. 491–495, doi: 10.15439/2016F332.

[6] 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,” in *Advances in Intelligent Systems and Computing*, vol. 700, 2018, pp. 282–297.

[7] G. I. Ulumudin, A. Adiwijaya, and M. S. Mubarok, “A multilabel classification on topics of qur’anic verses in English translation using K-Nearest Neighbor method with Weighted TF-IDF,” *Journal of Physics: Conference Series*, vol. 1192, no. 1, p. 012026, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012026.

[8] R. A. Pane, M. S. Mubarok, N. S. Huda, and Adiwijaya, “A multi-lable classification on topics of Quranic verses in English translation using multinomial naive bayes,” in *2018 6th International Conference on Information and Communication Technology, ICoICT 2018*, May 2018, pp. 481–484, doi: 10.1109/ICoICT.2018.8528777.

[9] N. S. Huda, M. S. Mubarok, and Adiwijaya, “A multi-label classification on topics of quranic verses (english translation) using backpropagation neural network with stochastic gradient descent and adam optimizer,” in *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, Jul. 2019, pp. 1–5, doi: 10.1109/ICoICT.2019.8835362.

[10] M. E. Aktas and E. Akbas, “Text classification via network topology: a case study on the holy quran,” in *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, Dec. 2019, pp. 1557–1562, doi: 10.1109/ICMLA.2019.00257.

[11] M. M. Chowdhury and A. Rahman, “Predicting places of revelation of Quran’s verses,” in *2020 International Conference on Computing and Information Technology, ICCIT 2020*, Sep. 2020, pp. 1–6, doi: 10.1109/ICCIT-144147971.2020.9213772.

[12] B. Arkok and A. M. Zeki, “Classification of Quranic topics using ensemble learning,” in *Proceedings of the 8th International Conference on Computer and Communication Engineering, ICCCE 2021*, Jun. 2021, pp. 244–248, doi: 10.1109/ICCCE50029.2021.9467178.

[13] B. Arkok and A. M. Zeki, “Classification of Quranic topics using SMOTE technique,” in *International Conference of Modern Trends in ICT Industry: Towards the Excellence in the ICT Industries, MTICTI 2021*, Dec. 2021, pp. 1–6, doi: 10.1109/MTICTI53925.2021.9664774.

[14] R. Hidayat and S. Minati, “Comparative analysis of text mining classification algorithms for English and Indonesian Qur’an translation,” *IJID (International Journal on Informatics for Development)*, vol. 8, no. 1, p. 47, Jun. 2019, doi: 10.14421/ijid.2019.08108.

[15] S. J. Putra, Y. Sugiarti, G. Dimas, M. N. Gunawan, T. Sutabri, and A. Suryatno, “Document classification using Naïve Bayes for Indonesian translation of the Quran,” in *2019 7th International Conference on Cyber and IT Service Management, CITSM 2019*, Nov. 2019, pp. 1–4, doi: 10.1109/CITSM47753.2019.8965390.

[16] O. V. Putra, F. D. Elfianita, T. Harmini, A. Trisnani, and A. Nugroho, “iQurNet: a deep convolutional neural network for text classification on the Indonesian Holy Quran translation,” in *2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021*, Jan. 2022, pp. 86–91, doi: 10.1109/ISMODE53584.2022.9743132.

[17] 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, Jan. 2020, doi: 10.31436/iiumej.v21i1.1170.

[18] Y. Purwati, F. S. Utomo, N. Trinarsih, and H. Hidayatulloh, “Feature selection technique to improve the instances classification framework performance for Quran ontology,” *International Journal on Informatics Visualization*, vol. 7, no. 2, pp. 615–620, Jul. 2023, doi: 10.30630/joiv.7.2.1195.

[19] T. Tantisripreecha and N. Soonthornphisaj, “A novel term weighting scheme for imbalanced text classification,” *Informatica (Slovenia)*, vol. 46, no. 2, pp. 259–268, Jun. 2022, doi: 10.31449/inf.v46i2.3523.

[20] H. Arabi and M. Akbari, “Improving plagiarism detection in text document using hybrid weighted similarity,” *Expert Systems with Applications*, vol. 207, p. 118034, Nov. 2022, doi: 10.1016/j.eswa.2022.118034.

[21] J. Attieh and J. Tekli, “Supervised term-category feature weighting for improved text classification,” *Knowledge-Based Systems*, vol. 261, p. 110215, Feb. 2023, doi: 10.1016/j.knosys.2022.110215.

[22] M. N. Raida, Z. N. Sristy, N. Ulfat, S. M. A. Monisha, M. J. I. Mostafa, and M. N. Haque, “A study on classifying Stack Overflow questions based on difficulty by utilizing contextual features,” *Journal of Systems and Software*, vol. 208, p. 111884, Feb. 2024, doi: 10.1016/j.jss.2023.111884.

[23] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, “Text classification using novel term weighting scheme-based improved TF-IDF for internet media reports,” *Mathematical Problems in Engineering*, vol. 2021, pp. 1–30, Mar. 2021, doi: 10.1155/2021/6619088.

[24] D. J. Benkendorf, S. D. Schwartz, D. R. Cutler, and C. P. Hawkins, "Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models," *Ecological Modelling*, vol. 483, p. 110414, Sep. 2023, doi: 10.1016/j.ecolmodel.2023.110414.

[25] X. Cheng, K. Huang, Y. Zou, and S. Ma, "SleepEGAN: a GAN-enhanced ensemble deep learning model for imbalanced classification of sleep stages," *Biomedical Signal Processing and Control*, vol. 92, p. 106020, Jun. 2024, doi: 10.1016/j.bspc.2024.106020.

[26] Y. Li, J. Jin, H. Tao, Y. Xiao, J. Liang, and C. L. P. Chen, "Complemented subspace-based weighted collaborative representation model for imbalanced learning," *Applied Soft Computing*, vol. 153, p. 111319, Mar. 2024, doi: 10.1016/j.asoc.2024.111319.

[27] S. Tao, P. Peng, Y. Li, H. Sun, Q. Li, and H. Wang, "Supervised contrastive representation learning with tree-structured parzen estimator Bayesian optimization for imbalanced tabular data," *Expert Systems with Applications*, vol. 237, p. 121294, Mar. 2024, doi: 10.1016/j.eswa.2023.121294.

[28] X. Weng *et al.*, "A joint learning method for incomplete and imbalanced data in electronic health record based on generative adversarial networks," *Computers in Biology and Medicine*, vol. 168, p. 107687, Jan. 2024, doi: 10.1016/j.compbiomed.2023.107687.

[29] L. Zhu, Y. He, and D. Zhou, "A neural generative model for joint learning topics and topic-specific word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 1–15, Dec. 2020, doi: 10.1162/tacl_a_00326.

[30] I. Iacobacci, M. T. Pilehvar, and R. Navigli, "Embeddings for word sense disambiguation: an evaluation study," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, vol. 2, pp. 897–907, doi: 10.18653/v1/p16-1085.

[31] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 16, pp. E3635–E3644, Apr. 2018, doi: 10.1073/pnas.1720347115.

[32] S. Fu, D. Su, S. Li, S. Sun, and Y. Tian, "Linear-exponential loss incorporated deep learning for imbalanced classification," *ISA Transactions*, vol. 140, pp. 279–292, Sep. 2023, doi: 10.1016/j.isatra.2023.06.016.

[33] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing*, vol. 143, p. 110415, Aug. 2023, doi: 10.1016/j.asoc.2023.110415.

[34] J. Tang, Z. Hou, X. Yu, S. Fu, and Y. Tian, "Multi-view cost-sensitive kernel learning for imbalanced classification problem," *Neurocomputing*, vol. 552, p. 126562, Oct. 2023, doi: 10.1016/j.neucom.2023.126562.

[35] W. Chen, K. Yang, Z. Yu, and W. Zhang, "Double-kernel based class-specific broad learning system for multiclass imbalance learning," *Knowledge-Based Systems*, vol. 253, p. 109535, Oct. 2022, doi: 10.1016/j.knosys.2022.109535.

[36] A. Amalia, M. S. Lidya, A. Andrian, E. M. Zamzami, and S. M. Hardi, "OLCBot: dissemination of interactive information related to Indonesia's Omnibus Law with the implementation of fuzzy string matching algorithm and sastrawi stemmer," in *Proceeding - ELTICOM 2022: 6th International Conference on Electrical, Telecommunication and Computer Engineering 2022*, Nov. 2022, pp. 178–181, doi: 10.1109/ELTICOM57747.2022.10037966.

[37] S. Fahmi, L. Purnamawati, G. F. Shidik, M. Muljono, and A. Z. Fanani, "Sentiment analysis of student review in learning management system based on sastrawi stemmer and SVM-PSO," in *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, Sep. 2020, pp. 643–648, doi: 10.1109/iSemantic50169.2020.9234291.

[38] B. Siswanto and Y. Dani, "Sentiment analysis about Oximeter as COVID-19 detection tools on Twitter using Sastrawi Library," in *2021 8th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE 2021*, Sep. 2021, pp. 161–164, doi: 10.1109/ICITACEE53184.2021.9617216.

[39] J. Zhang and R. Xie, "Word2vec- powered algorithm for efficient retrieval of bill of quantities," in *Proceedings - 2023 International Conference on Image Processing and Computer Vision, IPCV 2023*, May 2023, pp. 108–111, doi: 10.1109/IPCV57033.2023.00027.

[40] A. G. Ayar, S. Aygun, M. Hassan Najafi, and M. Margala, "Word2HyperVec: from word embeddings to hypervectors for hyperdimensional computing," in *Proceedings of the ACM Great Lakes Symposium on VLSI, GLSVLSI*, Jun. 2024, pp. 355–356, doi: 10.1145/3649476.3658795.

[41] H. M. Waidyasooriya and M. Hariyama, "Performance evaluation of word2vec accelerators exploiting spatial and temporal parallelism on DDR/HBM-based FPGAs," *Journal of Supercomputing*, vol. 80, no. 12, pp. 17192–17211, Aug. 2024, doi: 10.1007/s11227-024-06120-x.

[42] J. Flisar and V. Podgorelec, "Identification of self-admitted technical debt using enhanced feature selection based on word embedding," *IEEE Access*, vol. 7, pp. 106475–106494, 2019, doi: 10.1109/ACCESS.2019.2933318.

[43] W. Tian, J. Li, and H. Li, "A method of feature selection based on word2vec in text categorization," in *Chinese Control Conference, CCC*, Jul. 2018, vol. 2018-July, pp. 9452–9455, doi: 10.23919/ChiCC.2018.8483345.

[44] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," Universiteit van Amsterdam, 2003.

[45] A. Fabijan, B. Polis, R. Fabijan, K. Zakrzewski, E. Nowosławska, and A. Zawadzka-Fabijan, "Artificial intelligence in scoliosis classification: an investigation of language-based models," *Journal of Personalized Medicine*, vol. 13, no. 12, p. 1695, Dec. 2023, doi: 10.3390/jpm13121695.

[46] G. Shao, L. Tang, and J. Liao, "Overselling overall map accuracy misinforms about research reliability," *Landscape Ecology*, vol. 34, no. 11, pp. 2487–2492, Nov. 2019, doi: 10.1007/s10980-019-00916-6.

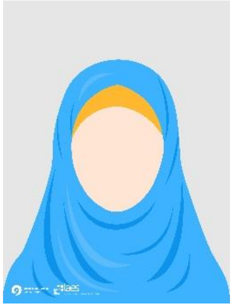
[47] G. Shao, H. Zhang, J. Shao, K. Woeste, and L. Tang, "Strengthening machine learning reproducibility for image classification," *Advances in Artificial Intelligence and Machine Learning*, vol. 2, no. 4, pp. 471–476, 2022, doi: 10.54364/AAIML.2022.1132.

[48] S. Kanj, F. Abdallah, T. Denœux, and K. Tout, "Editing training data for multi-label classification with the k-nearest neighbor rule," *Pattern Analysis and Applications*, vol. 19, no. 1, pp. 145–161, Feb. 2016, doi: 10.1007/s10044-015-0452-8.

[49] C. Loukas and N. P. Sgouros, "Multi-instance multi-label learning for surgical image annotation," *International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 16, no. 2, Apr. 2020, doi: 10.1002/rcs.2058.

BIOGRAPHIES OF AUTHORS

Dr. Fandy Setyo Utomo received his Master's in computer science at the Faculty of Mathematics and Natural Sciences, Gadjah Mada University in 2015. He received a Doctor of Philosophy (Ph.D.) from the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTEM) in 2023, specializing in information retrieval. Currently, he works as a lecturer at the Faculty of Computer Science, Universitas Amikom Purwokerto. His research areas are information retrieval and its usage in Quran and Islamic literature, text processing, machine learning, deep learning, and intelligent systems. He can be contacted at email: fandy_setyo_utomo@amikompurwokerto.ac.id.



Yuli Purwati received his Master's in computer from STMIK Amikom Yogyakarta in 2013. Currently, she is a senior lecturer at the Faculty of Computer Science, Universitas Amikom Purwokerto. Her research areas are software engineering, database systems, and distributed systems. She can be contacted at email: yulipurwati@amikompurwokerto.ac.id.

Dr. Mohd Sanusi Azmi is an associate professor at the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. He received a Doctor of Philosophy (Ph.D.) in computer science from University Kebangsaan Malaysia. He is the Malaysian pioneer researcher in the identification and verification of digital images of Al-Quran Mushaf. He is also involved in Digital Jawi Paleography. He actively contributes to the feature extraction domain. He has proposed a novel technique based on geometry feature used in Digit and Arabic based handwritten documents. His Fields of specialization are Feature extraction, handwritten character recognition, document analysis, digital paleography, and digital authentication of Al-Quran Mushaf. He can be contacted at email: sanusi@utem.edu.my.

Lulu Shafira is an active undergraduate student in the Informatics Department, Faculty of Computer Science, Universitas Amikom Purwokerto. Her research areas are software engineering and intelligent systems. She can be contacted at email: lshafira7@gmail.com.

Nikmah Trinarsih is an active undergraduate student in the Informatics Department, Faculty of Computer Science, Universitas Amikom Purwokerto. Her research areas are software engineering and intelligent systems. She can be contacted at email: nikmahtrinarsih02@gmail.com.