The Multi-Class Classification for the First Six Surats of the Holy Quran

Nouh Sabri Elmitwally[1], Ahmed Alsayat[2]
Department of Computer Science, College of Computer and Information Sciences, Jouf University, Aljouf, Saudi Arabia[1, 2]
Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt[1]

***Abstract*—The Holy Quran is one of the holy books revealed to the prophet Muhammad in the form of separate verses. These verses were written on tree leaves, stones, and bones during his life; as such, they were not arranged or grouped into one book until later. There is no intelligent system that is able to distinguish the verses of Quran chapters automatically. Accordingly, in this study we propose a model that can recognize and categorize Quran verses automatically and conclusion the essential features through Quran chapters classification for the first six Surat of the Holy Quran chapters, based on machine learning techniques. The classification of the Quran verses into chapters using machine learning classifiers is considered an intelligent task. Classification algorithms like Naïve Bayes, SVM, KNN, and decision tree J48 help to classify texts into categories or classes. The target of this research is using machine learning algorithms for the text classification of the Holy Quran verses. As the Quran texts consists of 114 chapters, we are only working with the first six chapters. In this paper, we build a multi-class classification model for the chapter names of the Quranic verses using Support Vector Classifier (SVC) and GaussianNB. The results show the best overall accuracy is 80% for the SVC and 60% for the Gaussian Naïve Bayes.**

Keywords—Text classification; machine learning; natural language processing; text pre-processing; feature selection; data mining; Holy Quran

I. Introduction

Text classification of the Holy Quran is a research topic researchers should pay attention to in the context of machine learning algorithms.

The Holy Quran is a book that was sent down from the heavens into the heart of the prophet Muhammad to be delivered to all human beings, not only Muslims. The sacred words were revealed by Allah and written into a meaningful textual format that could be analysed and classified using machine learning classification algorithms.

It is considered a comprehensive book covering every component of life and accessible to all people. It addresses the heart and mind as one.

The texts of the Holy Quran are fertile ground for natural-language processing and text classification. Their uniqueness and meanings distinguish the features. The Holy Quran is the first source of legislation in Islam. It is necessary to apply data-mining techniques to classify the verses into chapters (surats) intelligently based on machine learning techniques.

Furthermore, annotation of the verses of the Holy Quran's surats depends not only on the text itself but also on the ordering of the surats. Therefore, this study builds a model to classify and differentiate Quranic verses, according to their surats.

We have previously studied the architecture of the Arabic Language Sentiment Analysis (ALSA) [1]. We extended the concept of text classification to apply it to the Holy Quran's verses. The total number of verses in the Holy Quran is about 6000. Multi-class classification means that we need an automating model that enables classification of the texts accordingly. For this reason, this paper looks at the first six chapters from the Holy Quran; its approximately 1000 verses contain a total 8000 features for the training and testing data.

This paper is constructed as follows: the next section presents related work on multi-class text classification of the Holy Quran. Experimental method and analysis are covered in Section 3. Finally, the fourth section includes the results followed by the conclusions and anticipations of future work.

II. Related Work

The study detailed in [2] proposed an automation model that could classify Al-hadeeth features into Sahih, Hasan, Da'if, and Maudu, using machine learning techniques (LinearSVC, SGDClassifier, and LogisticRegression).

The author of [3] built a machine-learning model using an algorithm (KNN, SVM, and Naïve Bayes) classification model to annotate labels for the Quranic verses. The accuracy of the text-classification algorithms reached over 70% for the multi-labels of the Quranic verses.

The authors of [4] proposed a multi-label classification approach to the topics of Quranic verses using a k-Nearest Neighbor (KNN) algorithm with a weighted TF-IDF and TF-IDF.

Another research paper looked at the impact evaluation for four classification algorithms (SVM, KNN, Naïve Bayes and Decision Tree) to classify the topic of the Quranic Ayāts/verses [5]. The same concept as studied in [6] used the MultinomialNB classifier.

The authors of [7] used the Propbank Corpus to improve the performance of semantic argument classification on Quran data using the SVM Linear.

The authors of [8] applied the GBFS approach to label Quranic verses based on two major references, the commentary on the verses and the English translation. In addition, they proposed the IG-CFS technique to label Quranic verses of surats al-Baqara and al-Anaam [9].

III. Experiment and Analysis

The proposed model consists of four important phases as shown in the following framework architecture: 1) data collection, 2) text feature engineering, 3) The Term Frequency – Inverse Document Frequency (TF-IDF) feature representation, and 4) The GaussianNB and SVC classifiers. The framework architecture of the multi-class Quran framework classification is shown in Fig. 1.

A. *Data Pre-processing and Cleaning*

Before machine-learning modelling, we applied text pre-processing and cleaning techniques to extract features according to the following steps: remove the Arabic Tashkeel symbols (e.g., ◌ٍ◌ِ◌ُ◌َ◌ً); and remove consecutive Tatweel ('ـ') within Arabic characters.

B. *Corpus*

The corpus size was 954 verses collected from the first six surats of the Holy Quran. Table I shows generated descriptive statistics summarizing the central tendency, dispersion and the shape of the corpus' distribution.

Table II outlines the extracted sample from the Holy Quran corpus for the six classified categories ["Fatiha", "Albaqrah", "AlEimran", "Alnisaa", "Almayida", "Alaneam"] in the first column. The number of verses is shown in the second column. The selected verse and its translation appear in columns three and four.

C. *Exploratory Data Analysis*

The goal of the Exploratory Data Analysis (EDA) is to extrapolate on the breadth of information reflected by the corpus data. Fig. 2 shows the number of verses per corpus class.

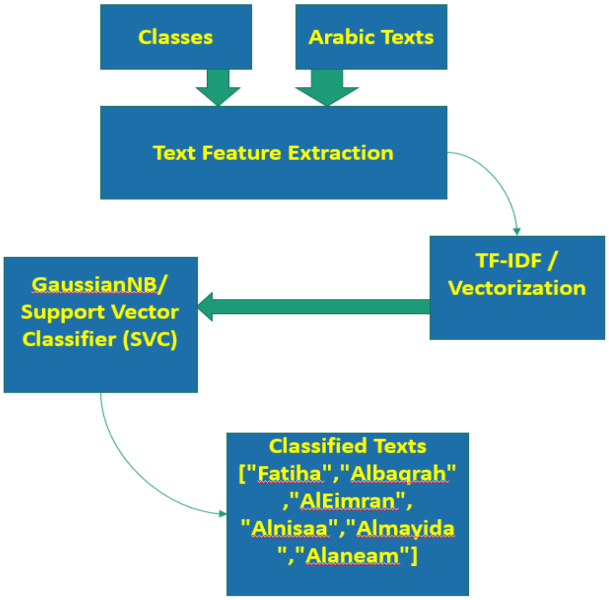
Classes
Arabic Texts
Text Feature Extraction
TF-IDF / Vectorization
GaussianNB/ Support Vector Classifier (SVC)
Classified Texts ["Fatiha","Albaqrah","AlEimran", "Alnisaa","Almayida","Alaneam"]

Fig. 1. The Quran Framework Classification.

TABLE I. The Descriptive Summary of the Holy Quran Corpus

count	954.000000
mean	3.640461
std	1.471548
min	1.000000
25%	2.000000
50%	3.000000
75%	5.000000
max	6.000000

TABLE II. Examples of Quran Verses

Chapter/Surat	Number of verses /Aya	Arabic Text	English Text
Fātiha/ the Opening Chapter/الفاتحة	7	بسم الله الرحمن الرحيم	In the name of God, Most Gracious, Most Merciful.
Baqara/ Heifer/البقرة	286	الم ذلك الكتاب لا ريب فيه هدى للمتقين	A.L.M This is the Book; in it is guidance sure, without doubt, to those who fear God
AlEimr Āl-i-'Imrānan/ The Family of 'Imrān/آل عمران	200	الم الله لا اله هو الحى القيوم	Allah! there is no god but He the Living the Self-Subsisting Eternal
Nisāa/The Women/النساء	176	يا أيها الناس اتقوا ربكم الذى خلقكم من نفس واحدة وخلق منها زوجها وبث منهما رجالا كثيرا ونساء واتقوا الله الذى تساءلون به والأرحام ان الله كان عليكم رقيبا	O mankind! reverence your Guardian-Lord Who created you from a single person created of like nature his mate and from them twain scattered (like seeds) countless men and women; reverence God through Whom ye demand your mutual (rights) and (reverence) the wombs (that bore you): for God ever watches over you.
Māida/ The Table Spread/المائدة	120	يا أيها الذين ءامنوا أوفوا بالعقود أحلت لكم بهيمة الأنعام إلا ما يتلى عليكم غير محلى الصيد وأنتم حرم إن الله يحكم ما يريد	O ye who believe! fulfil (all) obligations. Lawful unto you (for food) are all four-footed animals with the exceptions named: but animals of the chase are forbidden while ye are in the Sacred Precincts or in pilgrim garb: for God doth command according to His Will and Plan.
An'ām/Catle/الأنعام	165	الحمد لله الذى خلق السماوات والأرض وجعل الظلمات والنور ثم الذين كفروا بربهم يعدلون	

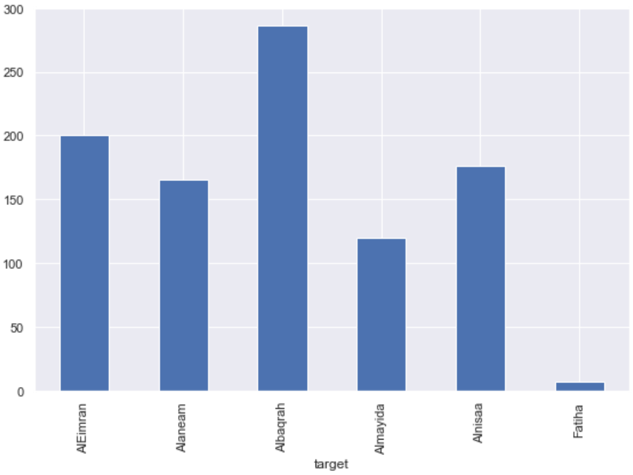

Fig. 2. Number of Verses Per Class.

D. Feature Engineering and Selection

Feature-text selection and engineering are considered the process of choosing the essential features required to represent the model for machine-learning classifiers. The following figures (Fig. 3-8) show word clouds for each Surat in the Holy Quran corpus.

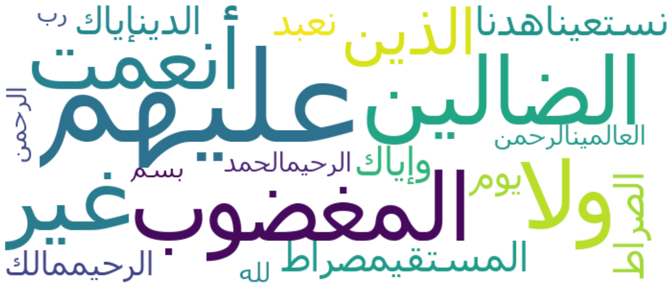

Fig. 3. Word Cloud-Surat Al Fatiha.

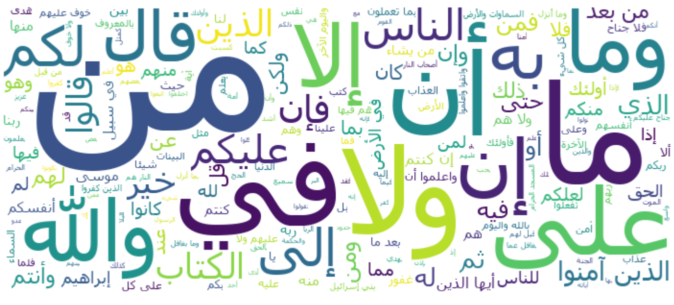

Fig. 4. Word Cloud-Surat Albaqara.

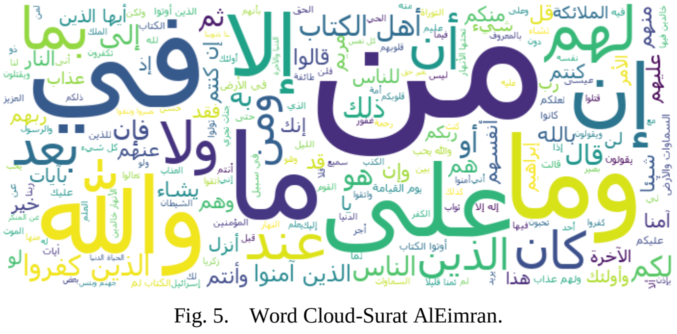

Fig. 5. Word Cloud-Surat AlEimran.

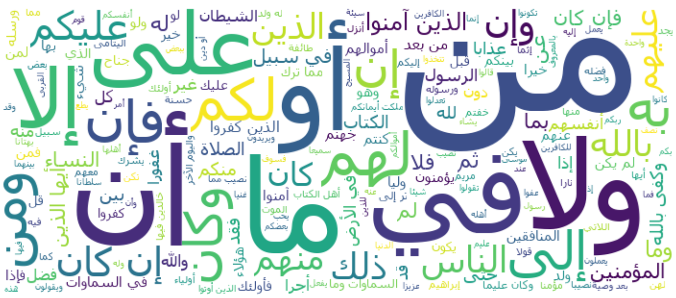

Fig. 6. Word Cloud-Surat AlNisaa.

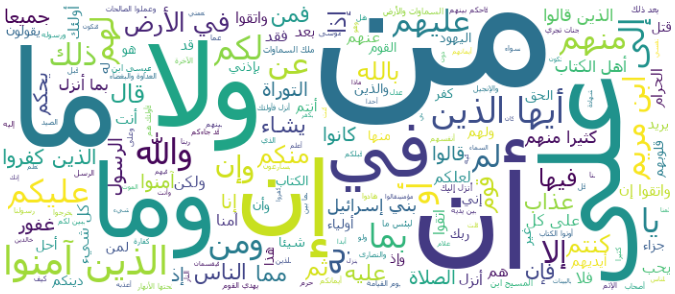

Fig. 7. Word Cloud-Surat AlMaeda.

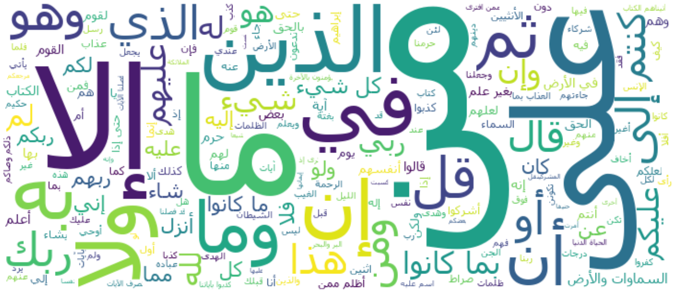

Fig. 8. Word Cloud-Surat Alanaam.

IV. Results

We calculated Accuracy, Recall and F1-value according to the following mathematical equations:

$$F - value = \frac{2*Accuracy*Recall}{Accuracy+Recall} * 100\% \qquad (1)$$

$$Recall = \frac{The\ number\ of\ corrected\ texts\ of\ a\ specific\ class}{The\ number\ of\ texts\ of\ this\ class\ in\ testing\ data} * 100\% \qquad (2)$$

$$Accuracy = \frac{The\ number\ of\ corrected\ texts\ in\ a\ specific\ class}{The\ number\ of\ texts\ in\ the\ class} * 100\% \qquad (3)$$

A. Machine-Learning Classifiers

The Support Vector Classifier (SVC) is considered the implementation of the Support Vector Machine (SVM) [5] for solving multi-class classification problems. The GaussianNB performs accurate feature-vector classification for the multi-class text problems [10]. We tested the proposed model against the performance metrics. The results are shown in Table III.

The sample texts of misclassified instance-classes are listed in Table IV. The table shows the missed classified text according to the expected and predicted output for the six classes (“Fatiha”–1; “Albaqrah”–2; “AlEimran”–3; “Alnisaa”– 4; “Almayida”–5; “Alaneam”–6).

B. Evaluation Metrics

The classification algorithms need the performance metrics to measure the model accuracy and losses. Fig. 9 shows that most of the performance metrics we used to evaluate the proposed multi-class Quranic model. The performance metrics are: 1) cohen_kappa; 2) log_loss; 3) zero_one_loss; 4) hamming_loss; and 5) Mathews_corrcoef.

The proposed model is evaluated according to two classifiers, SVC [7] and GaussianNB, as shown in Table V and Table VI and the Fig. 10 and Fig. 11. The performance of the proposed model is measured in terms of accuracy, precision, recall, f-measure, AUC, and ROC curves. The SVC classifier had the highest AUC value of 0.97 while the GaussianNB had the AUC value of 0.82 (see Fig. 12 and Fig. 13).

TABLE. III. THE PERFORMANCE METRICS

Metric	**SVC**	**GaussianNB**
cohen_kappa_score	0.408	0.395
log_loss	0.000	16.456
zero_one_loss	0.450	0.476
hemming_loss	0.450	0.476
matthews_corrcoef	0.420	0.396

TABLE. IV. THE MISCLASSIFIED INSTANCE-CLASSES

Text	**Expected Output**	**Predicted Output**
وقاتلوهم حتى لا تكون فتنة ويكون الدين لله فإن انتهوا فلا عدوان إلا على الظالمين	2	6
يا أيها الذين آمنوا لا تأكلوا الربا أضعافا مضاعفة واتقوا الله لعلكم تفلحون	3	5
واتل عليهم نبأ ابني آدم بالحق إذ قربا قربانا فتقبل من أحدهما ولم يتقبل من الآخر قال لأقتلنك قال إنما يتقبل الله من المتقين	5	2
وإذا سمعوا ما أنزل إلى الرسول ترى أعينهم تفيض من الدمع مما عرفوا من الحق يقولون ربنا آمنا فاكتبنا مع الشاهدين	5	2
إن أول بيت وضع للناس للذي ببكة مباركا وهدى للعالمين	3	2

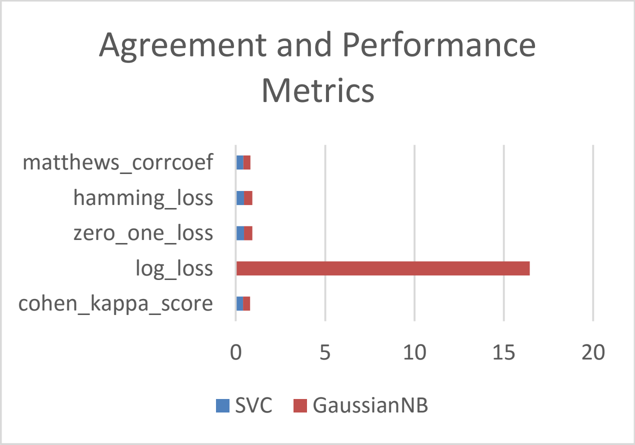

Fig. 9. The Agreement and Performance Metrics.

TABLE. V. RESULTS FOR SVM CLASSIFIER

Class	**Precision**	**Recall**	**F1- score**	**Area Under Curve (AUC)**
Fatiha	0.000	0.000	0.000	0.80
Albaqrah	0.487	0.475	0.481	0.68
AlEimran	0.545	0.364	0.436	0.85
Alnisaa	0.478	0.754	0.585	0.77
Almayida	0.871	0.771	0.818	0.97
Alaneam	0.444	0.167	0.242	0.76

TABLE. VI. RESULTS FOR GAUSSIANNB CLASSIFIER

Class	**Precision**	**Recall**	**F1- score**	**Area Under Curve (AUC)**
Fatiha	1.000	0.500	0.667	0.75
Albaqrah	0.424	0.350	0.384	0.61
AlEimran	0.548	0.515	0.531	0.71
Alnisaa	0.550	0.579	0.564	0.69
Almayida	0.686	0.686	0.686	0.82
Alaneam	0.355	0.458	0.400	0.67

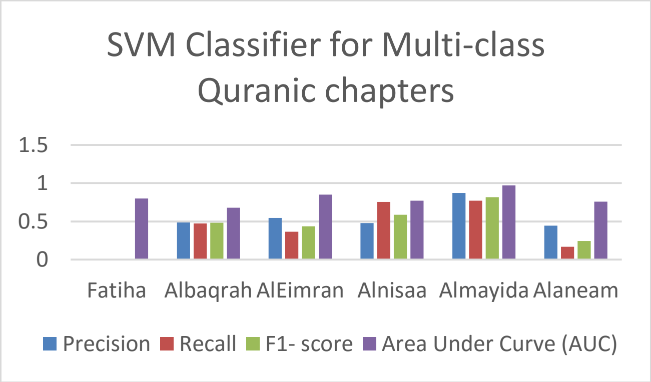

Fig. 10. SVM Classifier for Multi-Class Quranic Chapters.

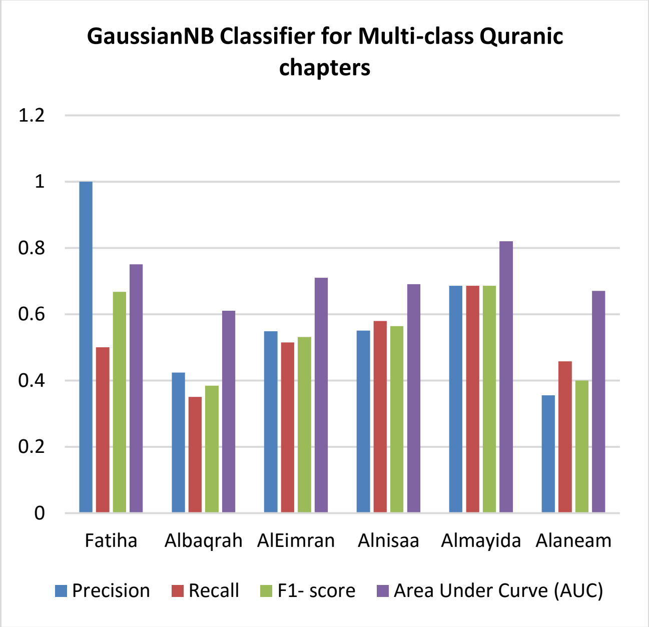

Fig. 11. GaussianNB Classifier for Multi-Class Quranic Chapters.

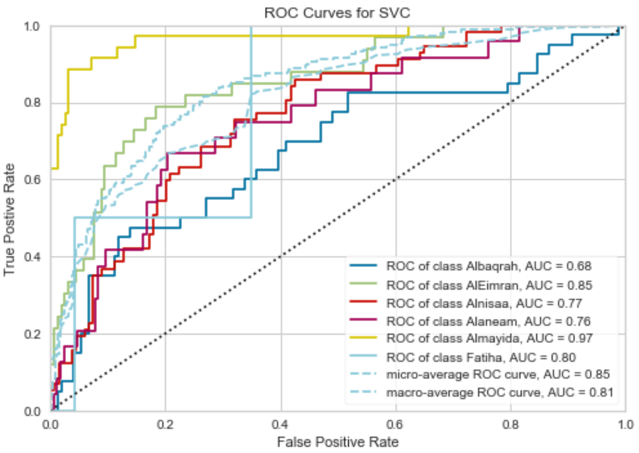

Fig. 12. ROC Curves – SVC.

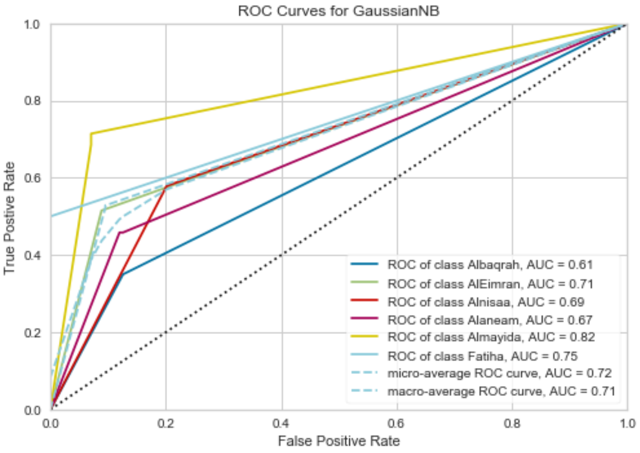

Fig. 13. ROC Curves – GaussianNB.

Finally, SVC [3] and GaussianNB classifiers were implemented for each verse of each Surat and measured the results in terms of the area under the curve (AUC) (see Fig. 14 and Fig. 15) [8]. The experimental results have shown that the proposed model had significant impacts on the multi-class Holy-Quran verse classification (see Fig. 16-19).

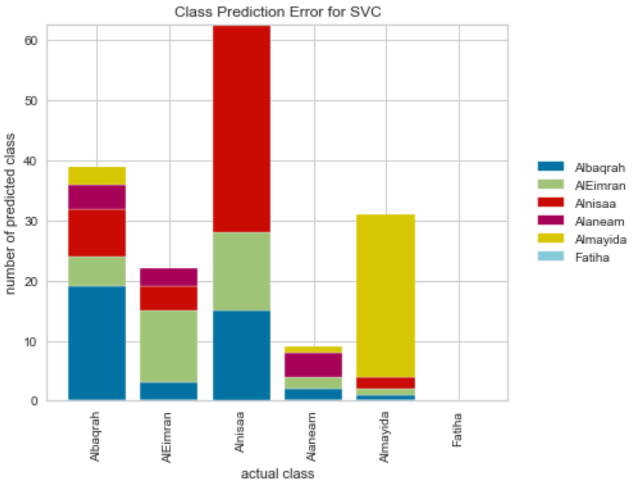

Fig. 14. Class Prediction Error for SVC.

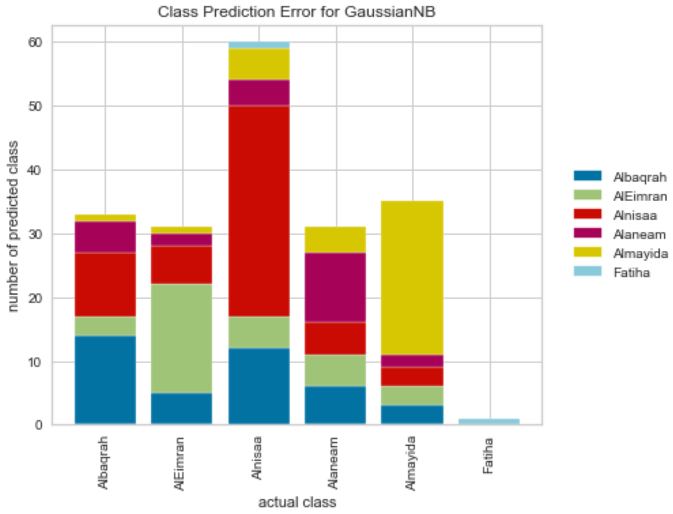

Fig. 15. Class Prediction error for GaussianNB.

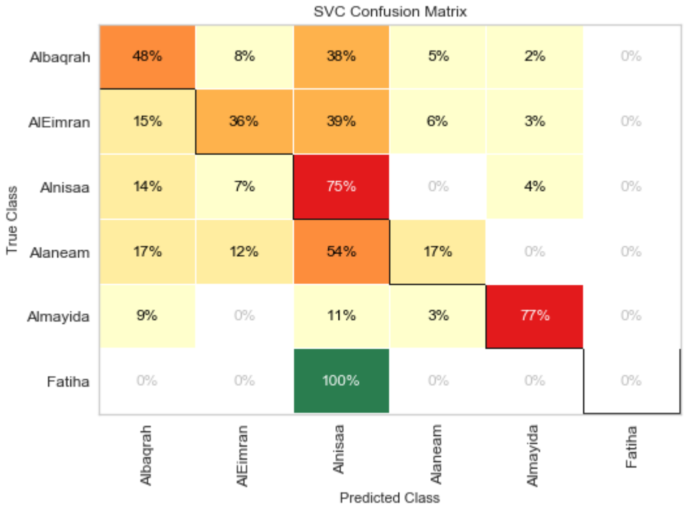

Fig. 16. SVC Confusion Matrix.

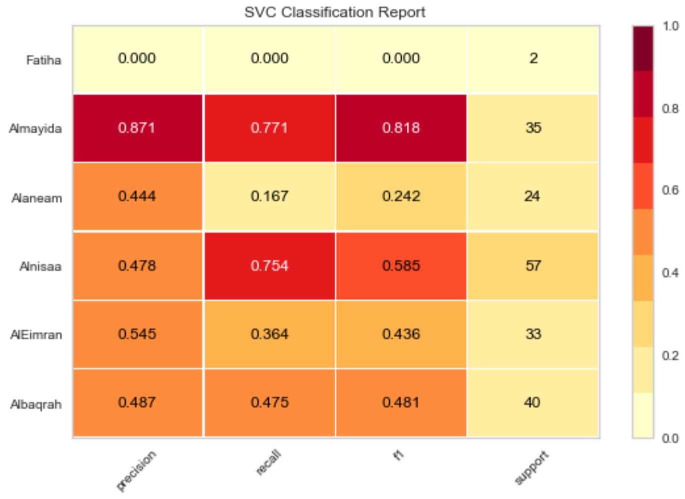

Fig. 17. SVC Classification Report.

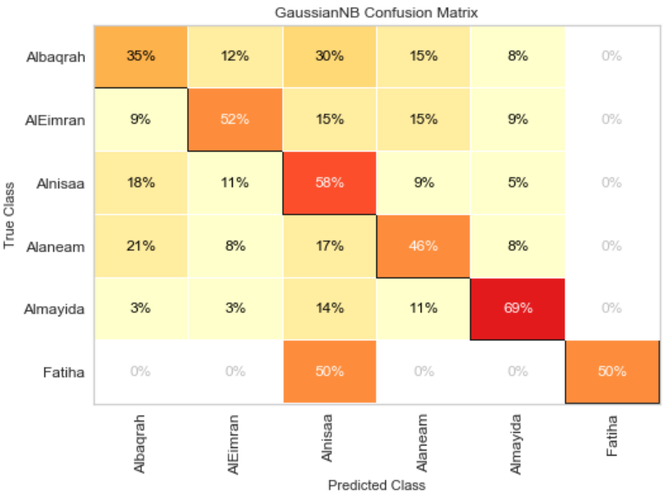

Fig. 18. Confusion Matrix—GaussianNB.

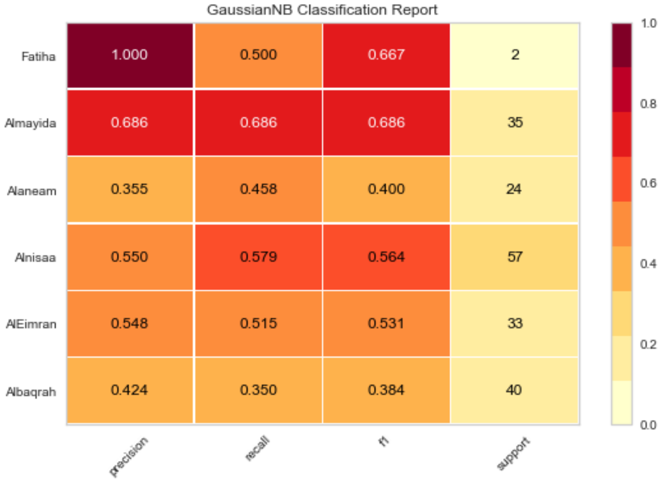

Fig. 19. Classification Report–GaussianNB.

V. Conclusions

Classifying chapters of the Holy Quran is considered a multi-class classification problem. In this paper, the multi-class classification for the Holy Quran corpus was used to train GaussianNB and SVC classifiers to predict the classification of the Quran verses into six surats. Increasing the size of the corpus and improved feature classification may improve the quality and accuracy of the framework. The experiment shows that the SVC provides the best results with an average of 88% f1-score. The research is to be continued by building a larger corpus for the verses of the Holy Quran chapters.

Acknowledgment

This research was fully funded by the Deanship of Scientific Research at Jouf University, AlJouf, Saudi Arabia on 2019/2020 using grant number (40/175); the APC was funded by Jouf University (JU).

References

[1] A. Alsayat and N. Elmitwally, “A comprehensive study for Arabic Sentiment Analysis (Challenges and Applications),” Egypt. Inform. J., p. S1110866519300945, Jun. 2019, doi: 10.1016/j.eij.2019.06.001.

[2] H. M. Abdelaal, A. M. Ahmed, W. Ghribi, and H. A. Youness Alansary, “Knowledge Discovery in the Hadith According to the Reliability and Memory of the Reporters Using Machine Learning Techniques,” IEEE Access, vol. 7, pp. 157741–157755, 2019, doi: 10.1109/ ACCESS.2019.2944118.

[3] A. O. Adeleke, N. A. Samsudin, A. Mustapha, and N. Nawi, “Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses.,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 4, p. 1419, Aug. 2017, doi: 10.18517/ijaseit.7.4.2198.

[4] 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,” J. Phys. Conf. Ser., vol. 1192, p. 012026, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012026.

[5] M. N. Al-Kabi, B. M. A. Ata, H. A. Wahsheh, and I. M. Alsmadi, “A Topical Classification of Quranic Arabic Text,” Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Its Sci., p. 7, 2013.

[6] 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), Bandung, 2018, pp. 481–484, doi: 10.1109/ICoICT.2018.8528777.

[7] D. K. Batubara, M. A. Bijaksana, and Adiwijaya, “On feature augmentation for semantic argument classification of the Quran English translation using support vector machine,” J. Phys. Conf. Ser., vol. 971, p. 012043, Mar. 2018, doi: 10.1088/1742-6596/971/1/012043.

[8] 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 Recent Advances on Soft Computing and Data Mining, vol. 700, R. Ghazali, M. M. Deris, N. M. Nawi, and J. H. Abawajy, Eds. Cham: Springer International Publishing, 2018, pp. 282–297.

[9] A. Adeleke and N. Samsudin, “A Hybrid Feature Selection Technique for Classification of Group-based Holy Quran Verses,” Int. J. Eng., p. 7.

[10] M. Karan, J. Šnajder, D. Sirinic, and G. Glavaš, “Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts,” in Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Berlin, Germany, 2016, pp. 12–21, doi: 10.18653/v1/W16-2102.