

Classification of Holy Quran Verses based on Imbalanced Learning

Bassam Arkok¹, and Akram M. Zeki²

^{1, 2} Kulliyyah of Information and Communication Technology, International Islamic University
Malaysia, Malaysia.

bassam_arkok@yahoo.com, akramzeki@iium.edu.my

Abstract

Imbalanced Learning (IL) is considered as a special case of text classification. It is applied in order to classify Imbalanced classes that are not equal in the number of samples. There are many researches on classified Quranic text which depends on different methods of classification. However, there is no study that classifies the Quranic topics based on Imbalanced Learning. So, this paper aims to apply the concept of IL to assign corresponding topics for the Quranic verses according to their contents. In this paper, two Quranic datasets have been classified by using Imbalanced Learning consecutively; the first dataset is Unification of God “Tawheed” and Polytheism of God “Shirk” verses, the second dataset is Meccan, and Medinan chapters. Imbalanced Classification is applied here since these topics have imbalanced classes which cannot be classified correctly by traditional methods. The results showed that applying Imbalanced Classification produces better outcomes than the results that are executed without using Imbalanced Classification techniques.

Keywords: Imbalanced Learning, Text Classification, Quranic Topics, Quranic Themes, Resampling, SMOTE.

1. Introduction

The Holy Quran is the Holiest Book of Allah SWT as it is considered as one of the main references for an estimated 1.6 billion Muslims around the world. There are many studies that have been conducted on the Holy Quran by using text classification techniques. Text classification is an important tool used today in many aspects of our life. Many studies have been performed to classify large-sized text documents using different standard of classifiers, ranging from simple distance classifiers such as K-Nearest-Neighbor (KNN) to more advanced classifiers such as Support Vector Machines (SVM). But the traditional approach fails when a short text is encountered because it is sparse due to the limited number of words.

Another common problem in text classification is class imbalance (CI). CI occurs when one class of the data contains most of the samples while the other class contains only a few samples. The standard classifiers, when applied to imbalanced data, results in high accuracy for the majority class and low accuracy for the minority class (Nayal, Jomaa et al. 2017)

Imbalanced text classification assigns one or more classes to a document according to their content for the text which is highly imbalanced. The data set is called class-imbalanced when one class consists of more samples significantly than the others and the imbalanced rates range is between 0.01 – 29.1% (Khalilia, Chakraborty et al., 2011). There are several methods used in solving the problem of classification for Imbalanced Classes, one of them is the oversampling method, which is one of the most common techniques discovered in solving this

problem. They are considered as pre-processing before the classification task that rely on generating new samples for the minority class that has few samples to rebalance or resample the class with the majority class that has more samples.

There are many imbalanced Quranic topics that cannot be classified by the traditional methods for example, Tawheed verses and Shirk verse, and also Meccan, and Medinan chapters. When these verses are classified, without taking into account the inequality in the number of samples, the overall accuracy will lead to the majority topic or class while the minority class will get misclassification. So, Imbalanced Learning is detected to solve these datasets that are used to classify many imbalanced datasets such as software defects, fraudulent credit card transactions, cancer gene expressions, telecommunications fraud, and natural disasters (Haixiang, Yijing et al., 2017).

There are many studies that classified Quranic verses for Arabic and non-Arabic language corpus to determine its corresponding topic. However, there is no research on classified Quranic topics based on IL. This research will classify imbalanced Quranic topics based on Imbalanced Text Classification techniques. The paper will be organized as follows: the next section is the literature review that will mention some studies based on classified Quranic datasets by different methods. Methodology of the research will be after that to clarify the steps and phases of the classification. Finally, the results and experiments will be explained to conclude some the findings and also to show results of the classification for the Holy Quran with or without using Imbalanced Learning.

2. Literature Review

There are many studies on classified Quranic verses for Arabic and non-Arabic language corpus to determine its corresponding topic. The first attempt of classifying Quranic Arabic topics was by (Al-Kabi, Kanaan et al., 2005). The authors designed a statistical model to categorize the verses automatically for the chapters of Fatiha and Yaseen for a number of predefined classes. A classifier classifies the verses in each chapter based on computing a score for every verse against each category as a first stage and then the verses were assigned to classes which have the highest score. This study is considered limited and primitive in the field of the topical classification of the Quranic verses (Al-Kabi, Alsmadi et al., 2013).

The authors (Al-Kabi et al., 2013) classified verses of the Quran according to its topics where they applied well-known classification algorithms: K-nearest neighbors (K-NN), decision trees, Naïve Bayes (NB), and support vector machines (SVM) for three selected topics only. Pre-processing phase was performed on the Quranic dataset and in addition, six performance metrics were used for the evaluation. However, only 1,227 verses were used out of the whole 6,236 verses of the Holy Quran for training and evaluating the selected topics. The topics were (Ignorant of religion, "الجاهلون بالدين"), (Oneness of God, "توحيد الله") and (Penalty of Apostates, "جزاء المرتدين"). These topics were considered based on the verses of the topics being unbalanced. There are 1,218 verses for Oneness of God, and 3 verses only for Ignorant of religion, while Penalty of Apostates has 6 verses. Therefore, the authors classified these imbalanced topics by using traditional classification which is been considered as the gap for this study, moreover, the evaluation of the performance for the classification of the imbalanced datasets was not used.

Akour et al., 2014, concentrated their research work on evaluating similarity of documents and text in Arabic language based on the Holy Quran. In this work, the Quranic verses were used as queries to search and evaluate similarity. The Measuring Quranic Verses Similarity and Sura Classification (MQVC) approach was employed for retrieving the most similar

verses in comparison with a user input verse as a query. The dataset consists of over 2000 verses from the Quran and they implemented the MQVC approach randomly on 70 of the 114 chapters of the Holy Quran. The experiment was performed by using N-gram technique besides Machine learning algorithm (LibSVM classifier) to classify the selected Quran chapters into chapters of Makki and Madani. The result showed the lowest precision value of 80% (for chapters 5, 13, 14, 63) and the highest precision value of 95% (for chapters 17, 32, 42).

Hassan et al., 2015, applied KNN algorithm to classify the Tafseer verses of the Holy Quran into predefined categories and to do this classification. A database of 1000 Ayat of the Quran was divided into two sets: training set consisting of 800 verses and test set consisting of 200 verses. Seven of the categories of the Tafseer texts were chosen which are: Inheritance, Marriage, Respecting Parents, Prayer, Zakat, Jihad, and Halal. After translating from Arabic to Malay language in which the results were evaluated based on Recall and Precision metrics. 'Inheritance' has the lowest recall value of 0.74 while 'marriage' category has the highest recall value of 0.9.

Hamed and Ab Aziz, 2016, aimed to develop the efficiency of information retrieval from the Holy Quran based on Question Answering System (QAS) through classifying the verses by using the Neural Network (NN) technique. This research had used the most popular English translation of the Quran of Abdullah Yusuf Ali as the data set. The QAS tackled some problems by expanding the question, using WordNet and benefitting from the collection of Islamic terms to avoid differences in terms of question and translations. In addition, this QAS classified the Al-Baqarah surah into two classes, which were fasting and pilgrimage based on the NN classifier, to reduce the retrieval of irrelevant verses since the questions of the user whereby asking about Fasting and Pilgrimage. Hence, the QAS retrieves the relevant verses to the question based on the N-gram technique, then ranking the retrieved verses based on the highest score of similarity to satisfy the desire of the user. By F-measure, the evaluation of classification by using NN had shown approximately 90% level and the evaluation of the proposed approach of this research based on the entire QAS has shown approximately 87% level.

Adeleke, Samsudin et al., 2017, presented a feature selection approach to label Quranic verses automatically so as to relate the verses with the five pillars of Islam; the 'Shahada' (profession of faith), the 'Salat' (daily prayers), the 'Zakat' (alms giving), the 'Saum' (fasting during Ramadan), and the 'Hajj' (pilgrimage to Mecca). The proposed feature selection approach was using group-based term count to represent the features extracted from three different Quranic translations. The approach adopted two of the common feature selection algorithms to reduce the feature space for 200 randomly selected verses based on the manual index, limited to the first two pillars of Islam. Finally, the k-NN, SVM, and NB classifiers were implemented independently on the feature selection algorithms to determine class membership for each verse and measure the results in terms of the area under the receiver operating characteristics curve (AUC). The experimental results had shown that the feature selection algorithms employed in the proposed approach had significant impacts on the classifiers implemented in the verse classification task. The lower the selected threshold, the more effective the features subset selection and the better the classifiers' performances.

Jamil, Ku-Mahamud et al., 2017, proposed a topic identification approach that classified Quranic verses translated to English in their appropriate topics. The proposed method used the term weighting scheme and it filtered the important terms from raw text to solve synonyms problem. In addition, it used a rule generation algorithm to identify the appropriate

topics that depends on the weighted terms. These topics were compared with topics identified by domain experts and using Rough Set. The findings show that the proposed approach was able to identify the topics closely to the topics given by the experts and Rough Set and it can be applied for different textual documents.

Jamil, Ku-Mahamud et al. 2017, proposed a topic identification method to identify subjects for groups of text that categorized Quranic verses to their topics based on the frequency technique. The algorithm of extraction that was presented to improve the results of extracted terms from the text and this algorithm were combined with the statistical method and computational linguistic method. The experimental results showed that the combined approach can be more effective for choosing relevant topic and outperformed the other extractions methods.

Adeleke, Samsudin et al., 2018, proposed a new feature selection algorithm called GBFS that is applied to classify Quranic verses based on the tafsir (commentary) and the English translation. The verses were from Al-Baqarah and Al-Anaam chapters for Faith, Etiquette, and Worship categories. The text data were preprocessed using StringToWord Vector with weighted TF-IDF. They implemented their experiments using feature selection algorithms: chi square, information gain, Pearson correlation coefficient, correlation-based, and relief which are compared with the proposed algorithm. They classified these topics by using four classifiers: LibSVM, naïve Bayes, KNN, Decision trees (J48).

The studies above will be summarized in Table 1 to show their limitations compared to the concept of Imbalanced classification.

Table 1. Previous studies compared to the concept of Imbalanced classification.

Reference	Classifier	Purpose	# of examples	IR	BC (Y/N)	IL (Y/N)
Nassourou, M., 2012.	SVM, NB	The Quran's chapters are classified into Meccan and Medinan by places of revelation.	- Meccan: 7 chapters. - Medinan: 7 chapters.	1	Y	N
Al-Kabi et al., 2013.	J48 ,KNN, SVM, NB	1227 verses are classified into Oneness of God, Penalty of Apostates, and Ignorant of religion.	- Oneness of God: 1,218 Verses. - Penalty of Apostates: 6 verses. - Ignorant of religion: 3 verses.	406:2:1	N	N
Akour et al., 2014.	LibSVM	The chapters are classified into Meccan, Medinan and Mix	- Meccan: 81 chapters. - Medinan: 21 chapters. - Mix: 12 chapters.	6.8:1 .8:1	N	N
Adeleke et al., 2017.	KNN, SVM, NB	200 verses are classified to Shahadah and Pray.	- Shahadah: 100 verses. - Pray: 100 verses.	1:1	Y	N

From the above Table, we can see that some studies used the traditional classification in which their datasets were imbalanced. Dataset without attention to their data is naturally imbalanced and the results are wrong because we cannot use the traditional methods and its

metrics with these data (Al-Kabi, Alsmadi et al., 2013, Akour, Alsmadi et al., 2014). Other studies classified their data as a balancing data without taking into account the data that is not imbalanced. (Nassourou, 2012). Moreover, after (Nassourou, 2012) classified their imbalanced data which is considered as balancing, he declare that the results and accuracy tend to one class, that is the majority class “Meccan”.

3. Research Method

In this section, methodology of the research will be explained in details. It comprises many steps before the classification task. Figure 1 reveals the flowchart of this work.

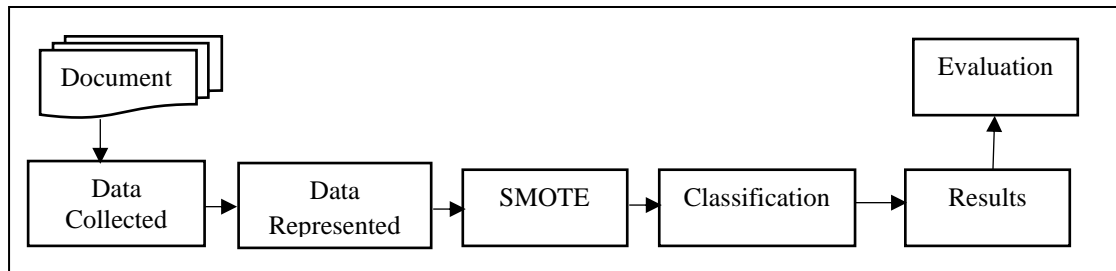


Figure 1. Flowchart of the research

3.1 Dataset Collected

The Quranic roots are already being extracted (Khedher, 2017) and used in this paper, so pre-presentation steps to get the words roots is cancelled and ignored. This work classifies verses of Tawheed and Shirk which are declared by the lexicon for the Quran’s topics that is used by (Al-Kabi et al., 2013) to make their experiments and get the results. The collected verses for the Tawheed topic are 962 verses while the verses for the Shirk verses are 118 where these topics are imbalanced that cannot be classified by the traditional methods correctly. In addition, Meccan, and Medinan chapters are imbalanced topics too because Meccan chapters are 86 chapters while the Medinan chapters are 28 chapters.

3.2 Data Representation

The collected data were represented using TF (term frequency). Firstly, all roots of the Quranic words were formatted in the vector space for matrix of the representation, after that, frequency of the words for each verse was counted and its value inputted under its feature in the space vector as all of these vectors are collected for all verses to represent the matrix for the classification phase. The vectors for verses of the Tawheed were labeled “Tawheed” while Shirk verses were labeled “Shirk” in the matrix. This policy of representation will be applied for Meccan, and Medinan chapters too.

3.3 SMOTE

The most common technique of oversampling (Blagus and Lusa, 2013, Skryjomski and Krawczyk 2017) is applied in this research. Synthetic Minority Over-sampling Technique (SMOTE) is proposed by (Chawla, et al., 2002) to oversample the minority class. The original samples for the minority class are used as a starting point to increase the minority class by generating artificial samples to rebalance the samples between the classes.

3.4 Classification phase

This work uses WEKA 3.9.2 (Waikato Environment for Knowledge Analysis) tool to upload dataset, in addition to implement SMOTE and the required classifiers. The most known classifiers are applied here, namely KNN, LibSVM, Random Forest, Decision Tree (J48),

Naïve Bayes, and Bayes Net. Before executing these classifiers, SMOTE - is considered as a pre-preprocessing step – and should be applied to oversample Shirk verses and Meccan chapters because they are the minority classes. This resampling makes rebalance for the imbalanced classes as Cross-validation; 10 folds is used in this work to get the results and to evaluate it.

4. Experiment and Results

The experiment and results will be for two datasets of the Holy Quran, the first dataset is for classification of Tawheed and Shirk verses and the second is for classification of Meccan, Medinan chapters. Imbalanced classification will be applied here because these verses have imbalanced classes that cannot be categorized by the traditional classification correctly.

4.1 Measures Extracted

This section mentions some of the metrics used for the imbalanced classification. (Haixiang, Yijing et al., 2017) mentioned that some metrics were used most frequently to evaluate performance of the imbalanced classification, namely, Sensitivity, Specificity, Overall Accuracy, Precision, Balanced Accuracy, F-Measure, G-mean, and Matthews Correlation Coefficient (MCC).

1. Sensitivity (also called the True Positive Rate, accuracy of positive examples, or recall): measures proportion of the positive examples that are identified correctly.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \dots \dots \dots \text{Equation. (1)}$$

2. Specificity (also called the True Negative Rate and accuracy of negative examples): measures proportion of the negatives examples that are identified correctly.

$$\text{Specificity} = \frac{TN}{TN + FP} \dots \dots \dots \text{Equation. (2)}$$

3. Overall Accuracy: It is a very essential parameter to test the efficiency of the model that is calculated as:

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \dots \dots \text{Equation. (3)}$$

In normal cases, accuracy alone provides adequate information to test precision of the system, but this seems not enough when dealing with imbalanced datasets, because a classifier may be over-fit and the prediction tends to be for the majority class (Tang, Zhang et al., 2009; Tang, Zhang et al., 2009; Al-Azani and El-Alfy, 2017; Patel and Thakur, 2017).

4. Precision (also called Positive Predictive Value): fraction of all the positive results that are truly positive examples.

$$\text{Precision} = \frac{TP}{TP + FP} \dots \dots \dots \text{Equation. (4)}$$

5. Balanced Accuracy: is defined as the arithmetic meaning of the specificity and sensitivity which avoids estimates of inflated performance for imbalanced datasets.

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity}}{\text{Specificity}} \dots \dots \dots \text{Equation. (5)}$$

6. F-measure (also called F1 score or F-score): this can be interpreted as a weighted average of the recall and precision measures which measures the importance trade-off between recall and precision.

$$F - \text{Measure} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \dots \dots \text{Equation. (6)}$$

7. G-mean: G-Mean is the geometric mean for specificity and recall. For binary classes, when accuracies of the classification are balanced, the accuracy will be maximized for each of them by G-mean.

$$G - \text{mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} \dots \dots \text{Equation. (7)}$$

8. MCC: Matthews correlation coefficient (MCC) takes the true and false positives and also the negatives into account. It is a correlation coefficient between the truly observed and the predicted binary classifications that it returns a value between -1 and $+1$. A coefficient of $+1$ indicates that the prediction is perfect while 0 indicates that the prediction is a worse prediction and -1 indicates total disagreement between the observation and prediction.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \dots \dots \text{Equation. (8)}$$

4.2 Results Discussion and Evaluation

The results and its evaluation are discussed and presented in this section. The results and experiments for Tawheed verses and Shirk verses will be presented first in 4.2.1 section while the results and experiments for the Meccan, and Medinan chapters will be shown in 4.2.2 section. The average of the metrics' values for each row and column is added in the Tables to provide us the general analysis for these metrics which will be discussed later.

4.2.1 Results of Tawheed and Shirk topics

This section presents the results and experiments for Tawheed and Shirk verses. Table 2 and the following figures [2-9] shows the results based on the metrics mentioned above. These results are revealed with and without applying SMOTE for the data to see the effect of imbalanced learning to improve performance of the classification for the imbalanced Quranic topics.

Table 2. reveals the results for Tawheed and Shirk verses

	Sensitivity	Specificity	Accuracy	Precision	Balanced Accuracy	F-Measure	G_Mean	MCC	AVG.
-SMOTE									
Bayes Net	0.97	0.40	0.91	0.93	0.69	0.95	0.62	0.44	0.74
Naive Bayes	0.87	0.51	0.83	0.94	0.69	0.90	0.67	0.31	0.72
LibSVM	1	0	0.89	0.89	0.50	0.94	0	?	0.60
KNN	0.91	0.23	0.84	0.91	0.57	0.91	0.46	0.14	0.62
J48	0.95	0.22	0.88	0.91	0.59	0.93	0.46	0.22	0.65

Random Forest	0.98	0.03	0.88	0.89	0.51	0.93	0.17	0.01	0.55
AVG.	0.95	0.23	0.87	0.91	0.59	0.93	0.4	0.22	
+SMOTE									
Bayes Net	1	0.88	0.94	0.89	0.94	0.94	0.94	0.89	0.93
Naive Bayes	0.91	0.69	0.80	0.74	0.80	0.82	0.79	0.61	0.77
LibSVM	0.88	0.79	0.83	0.80	0.84	0.84	0.83	0.67	0.81
KNN	0.61	0.96	0.78	0.93	0.79	0.74	0.77	0.61	0.77
J48	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.85	0.92
Random Forest	0.98	0.96	0.97	0.96	0.97	0.97	0.97	0.93	0.96
AVG.	0.89	0.87	0.88	0.88	0.88	0.87	0.87	0.76	

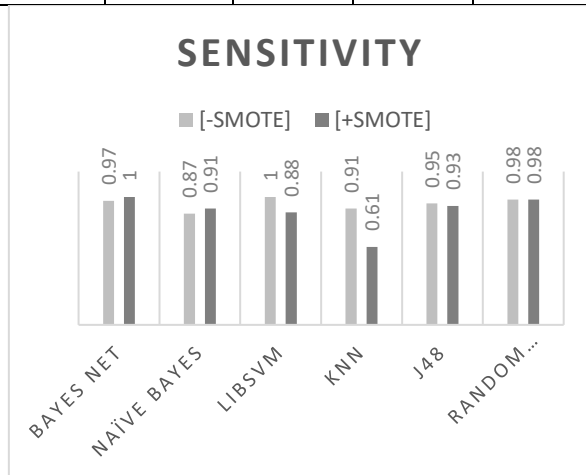


Figure 2. Measure of the Sensitivity for the Tawheed and Shirk topics.

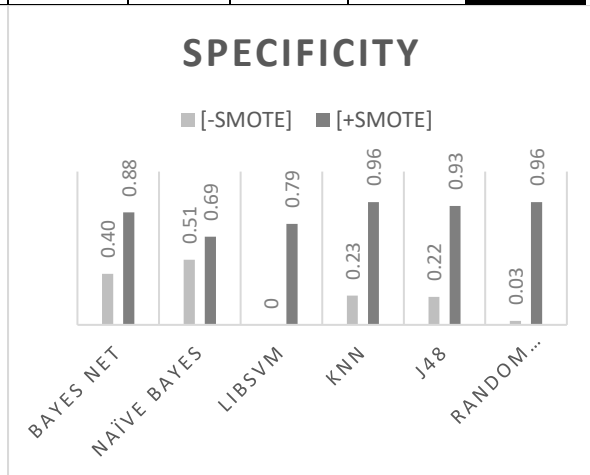


Figure 3. Measure of the Specificity for the Tawheed and Shirk topics.

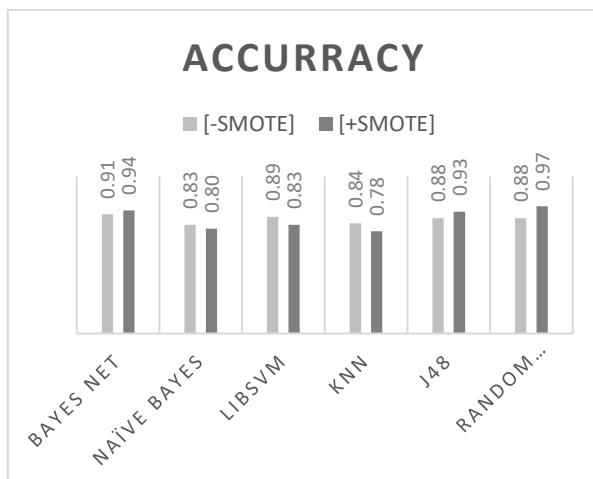


Figure 4. Measure of the Accuracy for the Tawheed and Shirk topics.

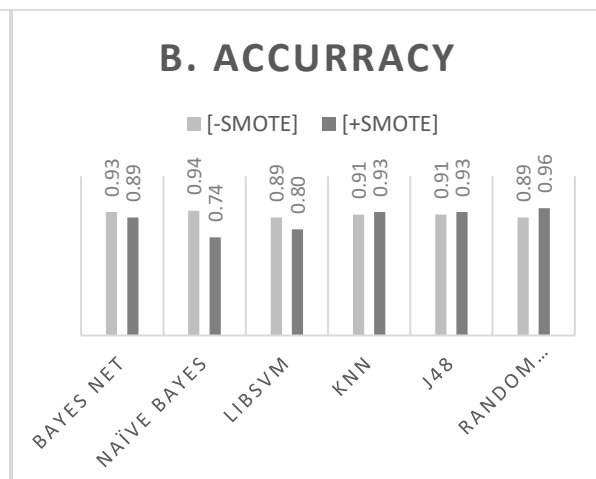


Figure 5. Measure of the Balanced Accuracy for the Tawheed and Shirk topics.

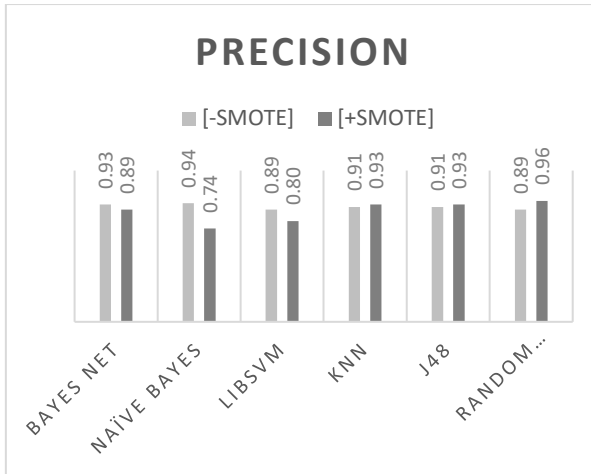


Figure 6. Measure of the Precision for the Tawheed and Shirk topics.

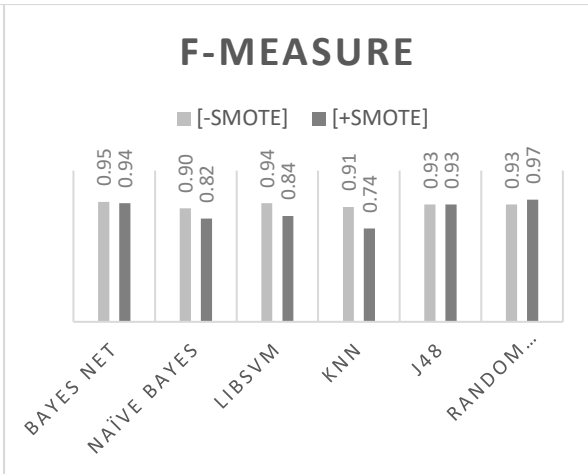


Figure 7. Measure of the F-Measure for the Tawheed and Shirk topics.

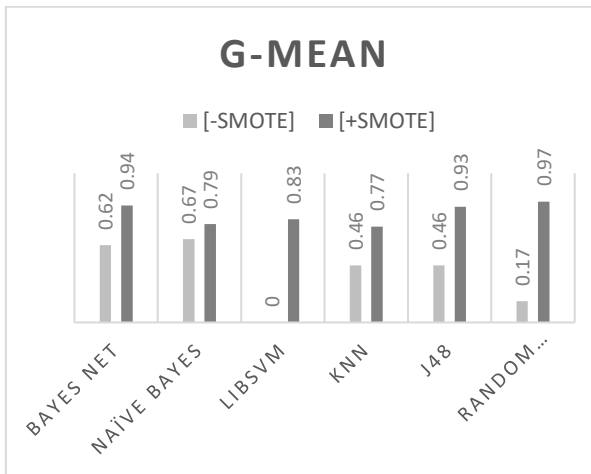


Figure 8. Measure of the G-Mean for the Tawheed and Shirk topics.

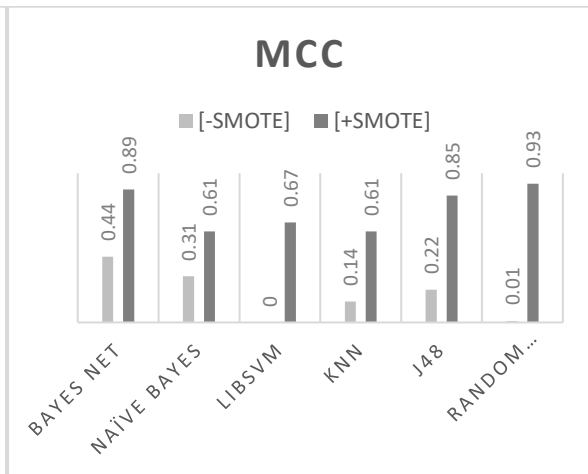


Figure 9. Measure of the MCC for the Tawheed and Shirk topics.

4.2.2 Results of Meccan, and Medinan chapters

This section presents the results and experiments for Meccan, and Medinan chapters. Table 3 and the following figures [10-17] reveal the results for these chapters according to the metrics above. These results are evaluated with and without applying SMOTE too as in the aforementioned topics.

Table 3 shows the results for Meccan, and Medinan chapters

	Sensitivity	Specificity	Accuracy	Precision	Balanced Accuracy	F-Measure	G_Mean	MCC	AVG.
-SMOTE									
Bayes Net	0.93	0.50	0.82	0.85	0.72	0.89	0.68	0.49	0.74
Naive Bayes	0.9	0.57	0.82	0.87	0.74	0.88	0.72	0.49	0.75
LibSVM	1	0.39	0.85	0.83	0.70	0.91	0.62	0.57	0.73
KNN	0.94	0.25	0.77	0.79	0.60	0.86	0.48	0.27	0.62
J48	0.88	0.50	0.79	0.84	0.69	0.86	0.66	0.41	0.70
Random	1	0.29	0.82	0.81	0.65	0.90	0.54	0.48	0.69

Forest									
AVG.	0.94	0.42	0.81	0.83	0.68	0.88	0.62	0.45	
+SMOTE									
Bayes Net	0.90	0.74	0.82	0.78	0.82	0.84	0.82	0.65	0.80
Naive Bayes	0.91	0.87	0.89	0.88	0.89	0.89	0.89	0.78	0.88
LibSVM	0.92	0.84	0.88	0.85	0.88	0.88	0.88	0.76	0.86
KNN	0.80	0.99	0.90	0.99	0.90	0.88	0.89	0.80	0.89
J48	0.91	0.88	0.90	0.89	0.90	0.90	0.89	0.79	0.88
Random Forest	1	0.94	0.97	0.95	0.97	0.97	0.97	0.94	0.96
AVG.	0.91	0.88	0.89	0.89	0.89	0.89	0.89	0.79	

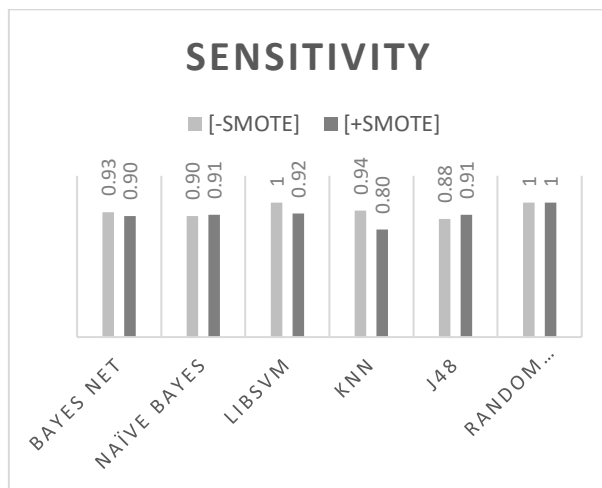


Figure 10. Measure of the Sensitivity for the chapters of the Meccan, and Medinan.

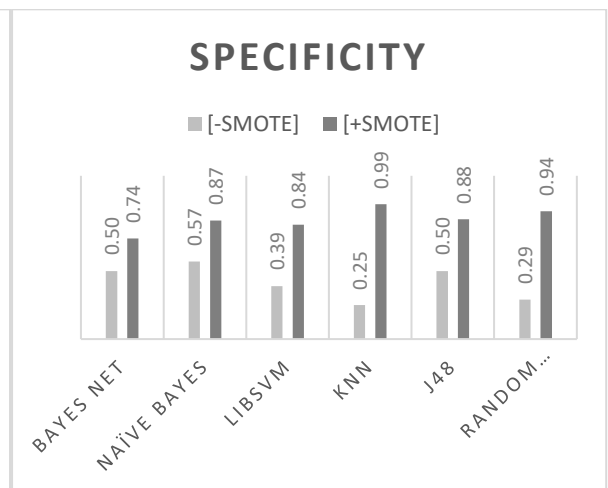


Figure 11. Measure of the Specificity for the chapters of the Meccan and Medinan.

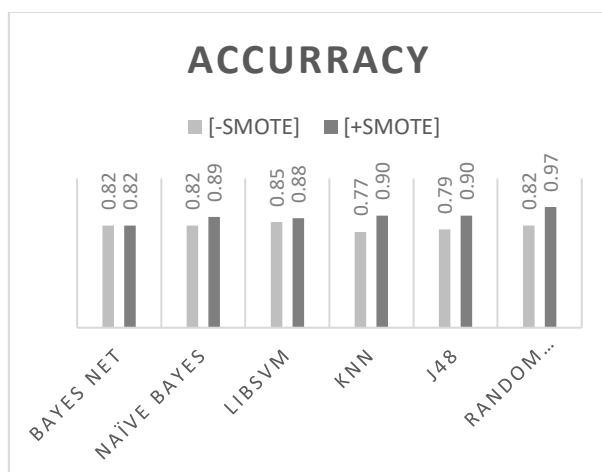


Figure 12. Measure of the Overall Accuracy for the chapters of the Meccan and Medinan.

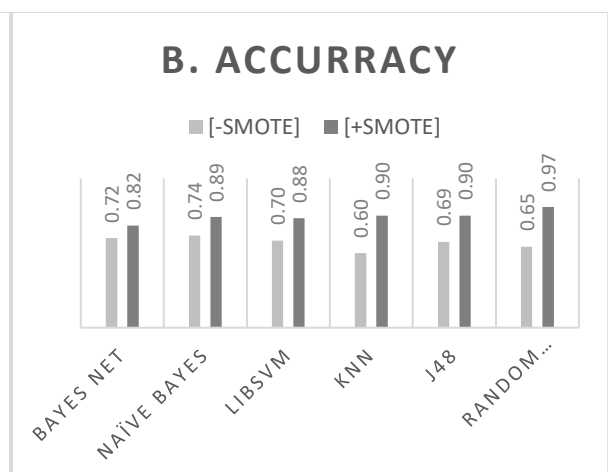


Figure 13. Measure of the Balanced Accuracy for the chapters of the Meccan and Medinan.

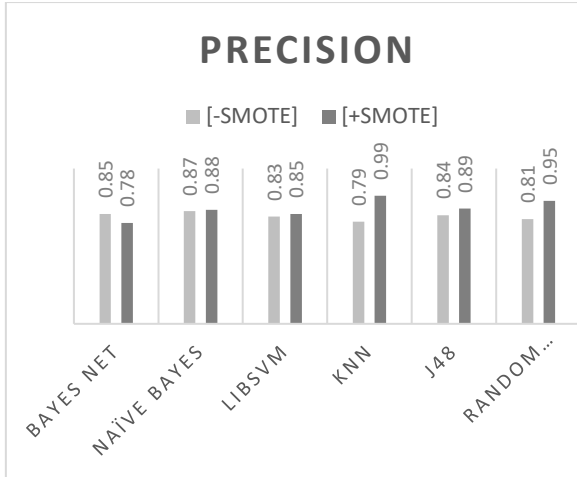


Figure 14. Measure of the Precision for the chapters of the Meccan and Medinan.

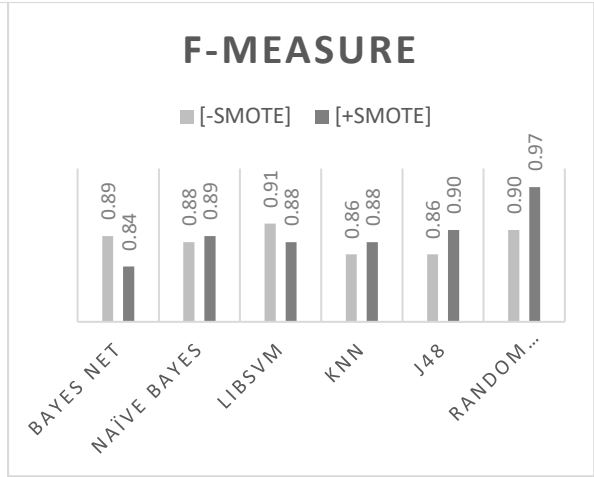


Figure 15. Measure of the F-Measure for the chapters of the Meccan and Medinan.

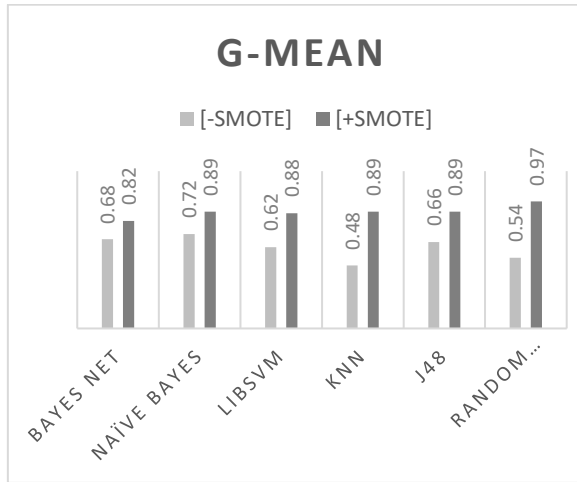


Figure 16. Measure of the G-Mean for the chapters of the Meccan and Medinan.

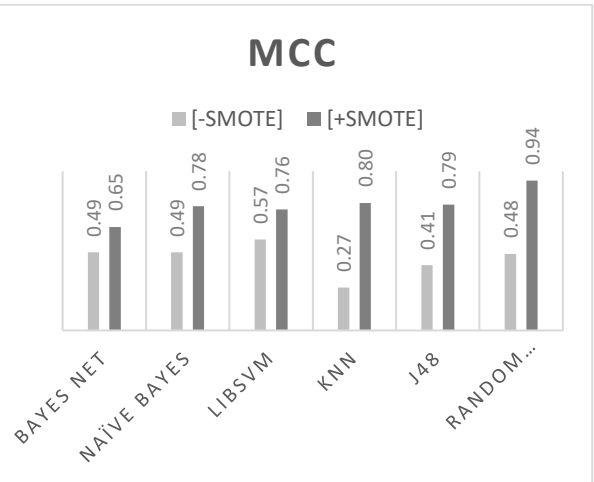


Figure 17. Measure of the MCC for the chapters of the Meccan and Medinan.

4.2.3 Analysis of the general results

Generally, and through the results obtained, performance of the classifiers for the Imbalanced Quranic topics are improved by using Imbalanced classification techniques with the stability of the overall accuracy. Also, all of the required classifiers categorized the datasets better after applying SMOTE as a pre-processing step before the classification phase. We can see that averages of the performance for all of the classifiers before SMOTE are in the range of 0.55 to 0.74 and 0.62 to 0.75 for the first dataset and the second dataset, respectively. Whilst this range improved after applying SMOTE for the datasets which becomes between 0.77 to 0.96 and 0.80 to 0.96 also for the first dataset and the second dataset, respectively.

Moreover, the classifiers can be ordered according to their performance from the strongest to the weakest before and after using Imbalanced Learning. Firstly, before applying SMOTE, the classifiers can be arranged as follow: Bayes Net, Naive Bayes, J48, KNN, LibSVM, and Random Forest for dataset of the Tawheed and Shirk. But for the dataset of the Meccan and Medinan chapters, the classifiers can be ordered as the following: Naive Bayes, Bayes Net, LibSVM, J48, Random Forest, and KNN. Secondly, after the implementation of SMOTE, the classifiers can be arranged as the following: Random Forest, Bayes Net, J48, LibSVM, KNN, and Naive Bayes for the dataset of the Tawheed and Shirk. While for the dataset of the

Meccan and Medinan chapters, the classifiers can be ordered as the following: Random Forest, KNN, J48, Naive Bayes, LibSVM, and Bayes Net.

We can also see that the averages of the evaluation for many of the metrics are affected and it improved significantly. For instance, the average of the Specificity metric for the dataset of Tawheed and Shirk was 0.23 before using Imbalanced Learning, but after the implementation of SMOTE, average of the Specificity became 0.87. This significant increase is acquired also for each of the metrics: Balanced Accuracy, G_Mean, and MCC as shown in the Table 2. However, some of metrics reduced very slightly after the implementation of SMOTE such as Sensitivity, Precision, and F-Measure. Similarly, for dataset of the Meccan and Medinan chapters, average of the Specificity metric increased from 0.42 to 0.88. There are many metrics which improved significantly also such as Balanced Accuracy, G_Mean, and MCC. In addition, Precision metric increased slightly while F-Measure is almost equal with the overall accuracy, whereas Sensitivity's value reduced slightly after applying SMOTE.

5. Conclusion

Text classification is one of the most important aspects used widely nowadays. There are several studies used in the Quranic text to classify it by applying different methods. However, there is no study for the classification of Quranic text based on Imbalanced Learning. So, this work aimed to classify imbalanced Quranic topics by using SMOTE to rebalance the dataset before the classification phase. Also, in this research, the performance of the classifiers is compared before and after using SMOTE with these datasets. The results showed that the performance of the classification improved significantly by applying Imbalanced Learning almost for many of the metrics with stability of the accuracy metric. The future work for this study is to apply the imbalanced learning for Sunnah too. Also, we can use other techniques for classifying imbalanced topics of the Quran and Sunnah to get better performance for these topics.

6. Acknowledgment


This paper was supported by International Islamic University of Malaysia under (FRGS19-083-0691) research project.

7. References

- Adeleke, A. O., et al. (2017). "Comparative analysis of text classification algorithms for automated labelling of Quranic verses." *Int. J. Advanc. Sci. Eng. Info. Tech* 7: 1419-1427.
- Adeleke, A. O., et al. (2018). A Group-Based Feature Selection Approach to Improve Classification of Holy Quran Verses. *International Conference on Soft Computing and Data Mining*, Springer.
- Akour, M., et al. (2014). "MQVC: measuring Quranic verses similarity and sura classification using N-gram." *WSEAS Transactions on Computers*.
- Al-Azani, S. and E.-S. M. El-Alfy (2017). "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text." *Procedia Computer Science* 109: 359-366.
- Al-Kabi, M. N., et al. (2013). A topical classification of Quranic Arabic text, *NOORIC 2013: Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*
- Al-Kabi, M. N., et al. (2005). "Statistical classifier of the holy Quran verses (Fatiha and Yaseen chapters)." *Journal of Applied Sciences* 5(3): 580-583.

- Blagus, R. and L. Lusa (2013). "SMOTE for high-dimensional class-imbalanced data." *BMC Bioinformatics* **14**(1): 106.
- Chawla, N. V., et al. (2002). "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* **16**: 321-357.
- Haixiang, G., et al. (2017). "Learning from class-imbalanced data: Review of methods and applications." *Expert Systems with Applications* **73**: 220-239.
- Hamed, S. K. and M. J. Ab Aziz (2016). "A question answering system on Holy Quran translation based on question expansion technique and Neural Network classification." *J. Comput. Sci* **12**(3): 169-177.
- Hassan, G. S., et al. (2015). "Categorization of 'Holy Quran-Tafseer' using K-Nearest Neighbor Algorithm." *International Journal of Computer Applications* **11**/2015 **129**(12): 1-6.
- Jamil, N. S., et al. (2017). "Topic Identification Method For Textual Document." *Journal of Multidisciplinary Engineering Science and Technology* .
- Jamil, N. S., et al. (2017). "A subject identification method based on term frequency technique." *International Journal of Advanced Computer Research* **7**(30): 103.
- Khalilia, M., et al. (2011). "Predicting disease risks from highly imbalanced data using random forest." *BMC medical informatics and decision making* **11**(1): 51.
- Khedher, M. Z. (2017). "Multiword corpus of the Holy Quran." *International Journal on Islamic Applications in Computer Science And Technology*. 2017 **5**(1).
- Nassourou, M. (2012). *Using Machine Learning Algorithms for Categorizing Quranic Chapters by Major Phases of Prophet Mohammad's Messengership*, Citeseer.
- Nayal, A., et al. (2017). "KerMinSVM for imbalanced datasets with a case study on arabic comics classification." *Engineering Applications of Artificial Intelligence* **59**: 159-169.
- Patel, H. and G. S. Thakur (2017). "Classification of imbalanced data using a modified fuzzy-neighbor weighted approach." *International Journal of Intelligent Engineering and Systems* **10**(1): 56-64.
- Skryjomski, P. and B. Krawczyk (2017). Influence of minority class instance types on SMOTE imbalanced data oversampling. *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*.
- Tang, Y., et al. (2009). "SVMs modeling for highly imbalanced classification." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(1): 281-288.

Biodata

	<p>Bassam Arkok is a PhD student at Faculty of Information and Communication Technology, International Islamic University Malaysia. His bachelor and master degree were in computer science. His research interest including Text mining, social networks analysis, Artificial intelligent algorithms and Machine learning, as well as the Islamic Applications in Computer Science and Technology.</p>
---	---



Akram M. Zeki is an Associate Professor at Faculty of Information and Communication Technology, International Islamic University Malaysia. His research interest including Watermarking, Information Security, Image Processing and Multimedia, as well as the Islamic Applications in Computer Science and Technology.

8. Abstract in Arabic

تصنيف لآيات قرآنية إعتماًداً على التصنيف الغير متوازي

بسام عركوك¹، أكرم محمد زكي²

^{2,1} كلية تقنية المعلومات والاتصالات، الجامعة الإسلامية العالمية الماليزية، ماليزيا

¹bassam_arkok@yahoo.com، ²akramzeki@iiium.edu.my

الخلاصة. التصنيف الغير متوازي أُعْتَبِر كحالة خاصة من تصنيف النصوص النصية والذي طُبِّق لتصنيف النصوص النصية الغير متساوية في عدد عيناتها. هناك باحثين صنفوا النصوص القرآنية والتي تعتمد على طرق مختلفة في التصنيف. ومع ذلك لا يوجد دراسة صنفَت النصوص القرآنية إعتماًداً على التصنيف الغير متوازي. لذا، هذا الورقة البحثية تهدف لتطبيق مفهوم التصنيف الغير متوازي لتعيين المواضيع المقابلة للآيات القرآنية وفقاً لمحتوهم. في هذه الورقة يوجد مجموعتين من المواضيع القرآنية صنفَت بإستخدام التصنيف الغير متوازي. المجموعة الاولى لآيات التوحيد والشرك، بينما المجموعة الثانية للصور المكية والمدنية. التصنيف الغير متوازي طُبِّق في هذا البحث لأن هذه المواضيع لديها مواضيع غير متوازية التي لا يُمكن تصنيفها بشكل صحيح بواسطة الطرق التقليدية. النتائج لهذا البحث أظهرت ان تطبيق التصنيف الغير متوازي كانت أفضل من النتائج التي نُفذت بدون إستخدام تقنيات التصنيف الغير متوازي.

الكلمات الجوهرية. التصنيف الغير متوازي، تصنيف النصوص النصية، المواضيع القرآنية، إعادة توزيع العينات،