# **POLS: 8500 Final Project**

# **Text Analysis in Arabic:**

# **Finding use of Quran Verses in Jihadi Texts**

# **By: Maulik Shah**

# **Institute of Artificial Intelligence,**

# **University of Georgia.**

**Introduction:**

It is noteworthy that in religious texts like Bible and Quran, there are some verses, which have relation with historic violence or advise to incur a certain type of violence to own self or others. Although, there can be many interpretation of each verse, some people around the world use the violent verses in the most provocative format to pursue their followers to commit violence. Having this scenario as a background, the project is aiming towards finding the use of violent verses from the religious text in a provocative form.

The project takes Quran as a reference religious text and find its use in the articles of Jihadis, where some of the Bin Laden’s documents are used as Jihadi documents reference. The project tries to find use of the Quran verses in Bin Laden’s documents while finding those verses, while it also tries to formulate a model from Quran texts to classify if a document can be termed as violent in terms of Quran verses.

For the exact text matching, Cosine Similarity and LDA Similarity methods are used, while for the creating of the predictive model, Logistic Regression, Naïve Bayes, and Support Vector Machine techniques are being explored. All the methods are explored in R, while the visualization of the results was also done in the R using some libraries available.

Although, there is very less literature available on the Arabic text matching, in the Mustafa, Suleiman H. (2005), they have tried to match and stem the Arabic texts using the N-gram model. Apart from that in the Al-Onaizan, Yaser, and Kevin Knight (2002), the authors try to transliterate the Arabic text. Current project uses “ArabicStemR” library by Rich Neilson, which can be said as an easy to use tool for Arabic text cleaning, stemming and transliterating. Muflikhah, Lailil, and Baharum Baharudin(2009) shows the use of Cosine similarity in text matching applications, which shows the benefit of this technique over Euclidian distance.

To find appropriate predictive model, there are three major techniques used as mentioned above. Thorsten(1998) has mentioned the use of Support Vector Machines in the text classification, Yiming(1999) has given the description of the use of Naïve Bayes in the text classification. Although, these are very basic approaches for text classification, the project currently focuses more on finding the similarities between the texts.

Notably, the project is an extension to Dr. L Jason’s previous research on finding the violence score for each verse in Quran and Bible.

**Data:**

The project has been designed to find the similarities between the Quran verses and the Jihadi texts. For the Quran verses, there are 6 different type of Quran versions available from different sources in Arabic. Apart from that as mentioned above, violence score for each verse is available from Dr. Jason’s previous researches. Apart from that some of the Bin Laden’s documents were available in the text and pdf formats.

Although the data was enough to start a research, it was in a crude format. To convert the data into an organized format, below steps were performed:

1. All the Quran versions from different text file were merged in a single csv file, with the data of violence score and English translation of each verse. The final CSV document was made up was of verses and surah numbers of Quran, violence score for each verse, all different Arabic versions of Quran and English translation of each verse. Here, a verse should be considered as a document.
2. From the available documents, randomly 12 documents were chosen to create a sample data set for analysis. The final document consisted of 575 different paragraphs of Bin Laden, with each indicating reference to its source document. Here, a paragraph should be considered as a document.

**Methods:**

This section can be divided into two parts, one is finding the most similar verses in Quran if referenced in the current texts and second is creating the predictive model for classifying violent documents.

1. Text Similarity between Jihadi Document and Quran Verses:

The primary task here is to find the most similar verse from the Quran, which is being referenced from the Jihadi texts. Two different methods were used for this purpose.

1. Cosine Similarity:

The method finds the distance between two vectors by finding the angle between each, while summing up the overall similarity between the texts. Cosine score 1 indicates a pure similarity, while 0 indicates the no similarity between two documents.

1. LDA Similarity:

LDA is an algorithm from the topic modelling family, which finds different topic used in a set of documents in an unsupervised way. The topic distribution in each document can then be used for finding similarities between the two documents.

1. Create Predictive Model:
2. Logistic Regression:

The logistic regression is a method, which is one of the primitive methods of the text classification. From the given inputs, this regression method tries to predict one of the two outputs mainly using log function. If the final log value of the function is below 0.5(can vary) then the document is of one class and of another class otherwise.

1. Naïve Bayes:

The Naïve Bayes is also one of the most widely used classification technique, which uses the probabilistic frequentist approach to derive the class of a document.

1. Support Vector Machines(SVM):

An SVM is a supervised machine learning method for classification, whose model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

**Implementation Procedure:**

This section is primarily divided into 4 sections, which are text cleaning and transformation, finding text similarity, predictive modelling, and statistics. All the techniques on the required data are implemented in R. The section also specifies the libraries used in terms for implementation.

1. Text Cleaning and Transformation:

The most important part in any text analysis is text cleaning and preparing it for analysis into a required format. As, the current text is in Arabic, each document is first loaded as a UTF-8 format in the R. For the text cleaning part, “ArabicStemR” library is used, which removes punctuation marks and, unnecessary prefix and postfix characters. The library is also used for stemming the words and transliterate the Arabic text into English in respective order.

All the techniques for text matching and similarity requires the text to be converted into a document – term matrix. The cleaned text is then used for creating a document-term matrix(DTM), which is done using the methods from a standard R library “e1071”.

1. Text Matching:

Both the text matching algorithms uses the document term matrix as their inputs to calculate the similarity. For finding the cosine similarity, “lsa” library is used. Each document is compared with each Quran verse to find the similarity scores and in the end one matrix is created with the maximum similarity scores between a document and a verse.

While for the LDA similarity, the “topicmodels” library is used, which first creates a topic model using Quran verses and then tries to predict the topics of the document by the newly created topic model. As in the Cosine Similarity procedure, each document topic model distribution is compared with topic model distribution of each Quran verse to find a score. The maximum score of this LDA procedure would give the required output.

1. Predictive Modelling:

As already mentioned all the predictive modeling techniques here would require a document term matrix, except Naïve Bayes. For the Naïve Bayes modelling, the DTM is converted into a frequency matrix.

All 3 methods Logistic Regression, Naïve Bayes and SVMs are being referenced from “e1071” package. As the Bin Laden documents were not classified, the data used in the prediction was only Quran verses. The Quran verses were randomly sampled 85% for training and 15% for the testing. The document term matrix was created after that to find the overall performance.

1. Statistics:

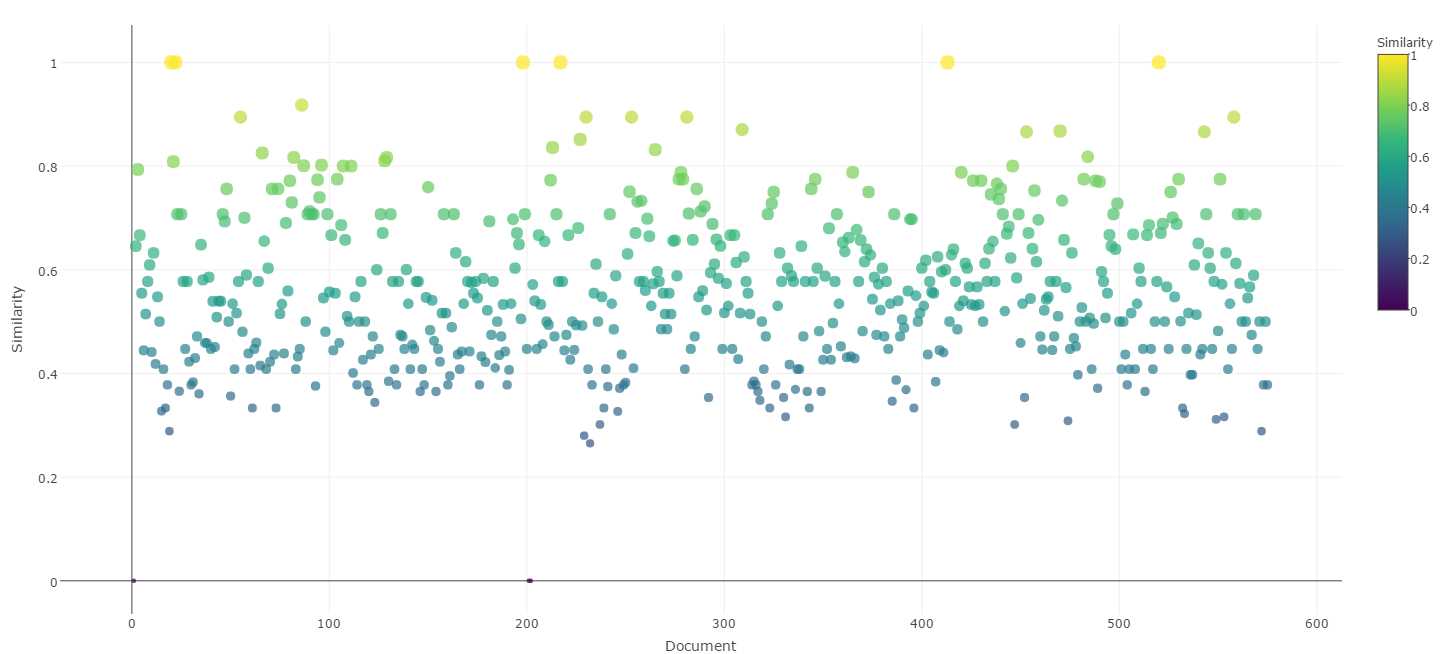
For finding the statistics of the results, mean and standard deviation of the similarity scores are calculated, while accuracy, specificity and sensitivity are the measure of the predictive models. The graphics used in this literature are created using “plotly” library.

**Results:**

1. Text Matching:

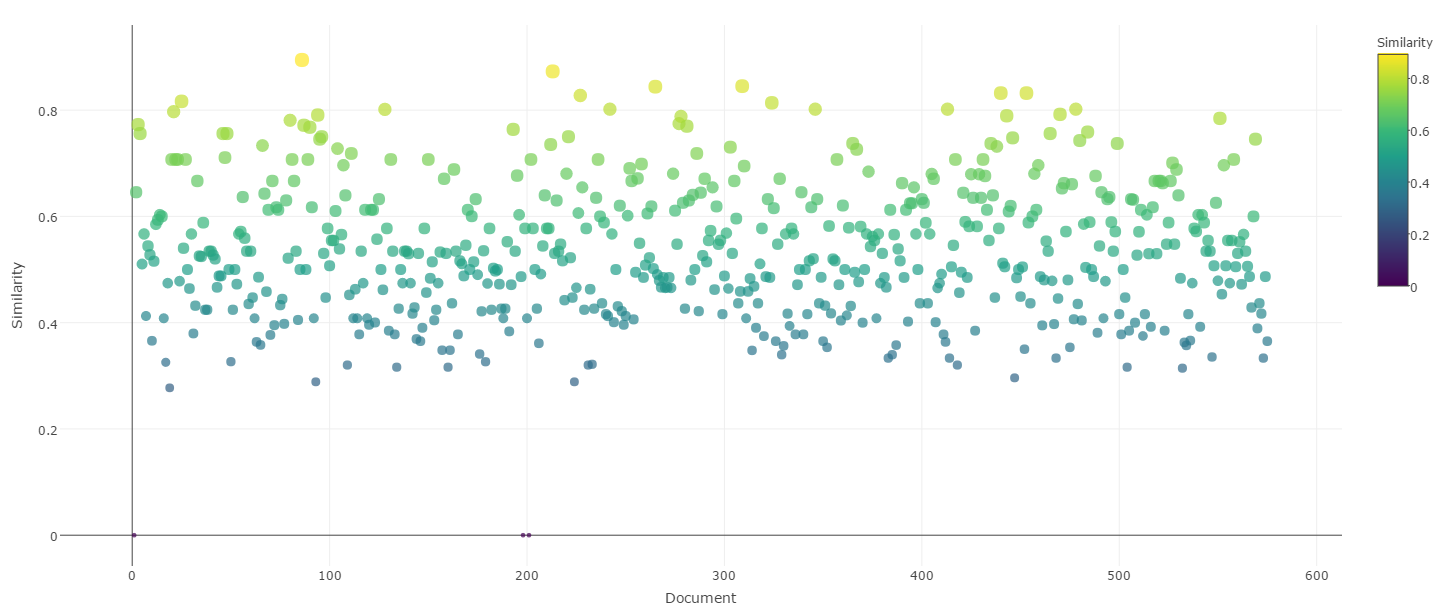
For finding the Cosine score and LDA similarity scores, DTMs with stemming and without stemming documents were created to find the performance. An external human Arabic expert found a threshold of 0.7 for the cosine similarity to be considered as a reliable score. Below plots show the overall results of both the similarity methods for the 575 documents(paragraphs) of Bin Laden:

*Cosine Similarity distribution with Text Stemming*



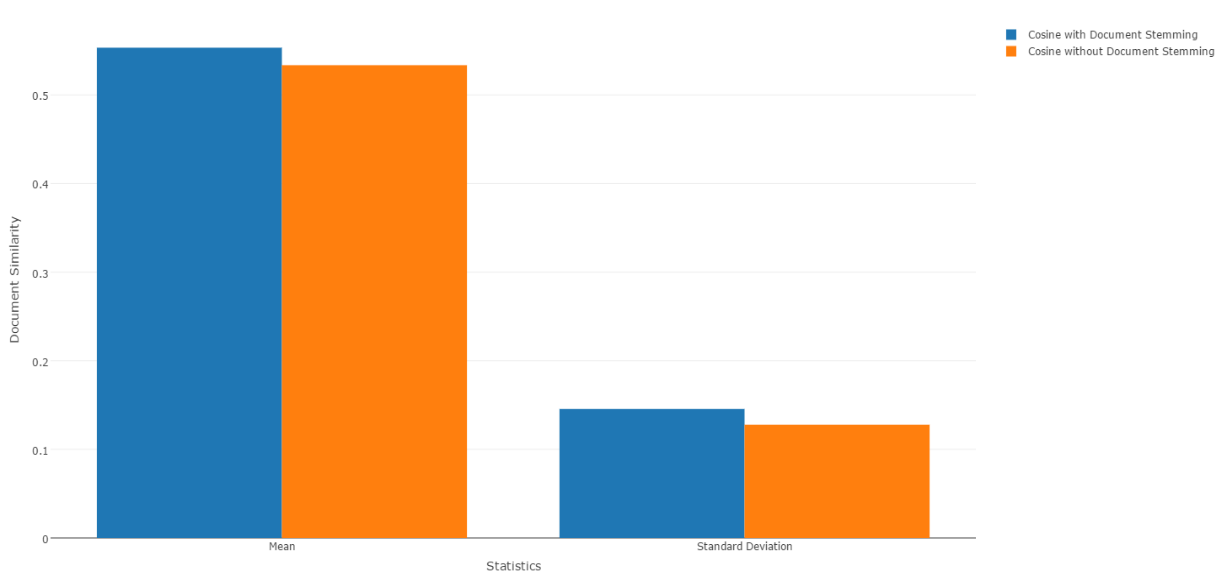
The graph indicates a high level of dissimilarity between Quran verses and Bin Laden’s documents, though the number of similar paragraphs are also significant. Although, the most significant finding was that the almost none of the verses used from Quran in the documents were violent.

*Cosine Similarity distribution with Text Stemming:*



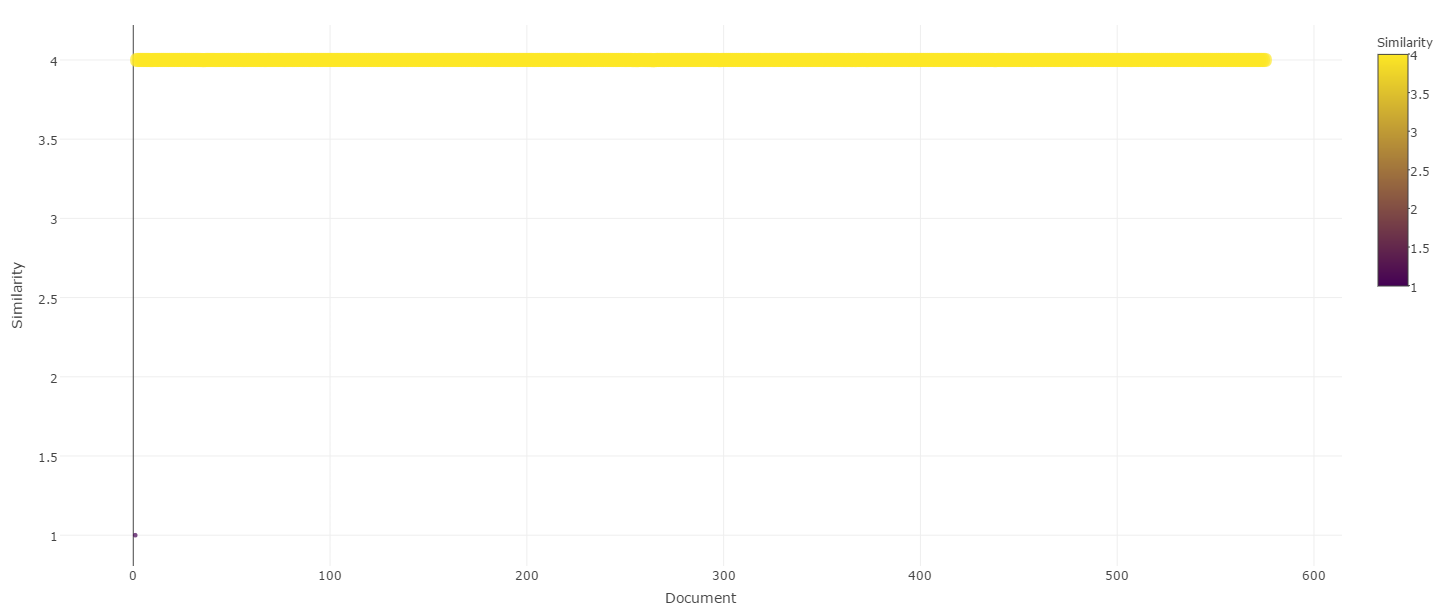
As shown in figure, the distribution is looks like as of the with stemming document, but the issue here is the verses are not the same in the output, which needs a little attention to the results.

Overall mean and standard deviation:



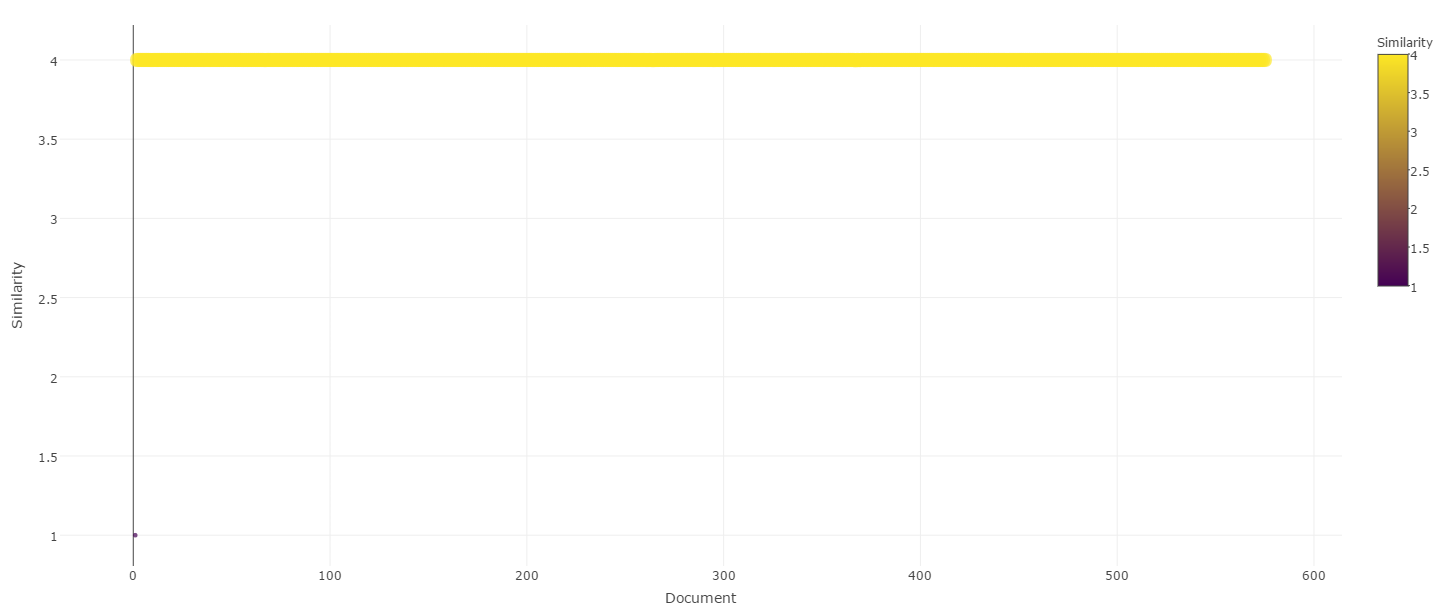
The mean of the score is near to 0.5 in both the cases where the with stemming scores are a bit more than their counter part in without stemming, while the document stemming has more standard deviation then the without stemming one which is about 0.125 and 0.12 respectively.

*LDA similarity with Stemming*



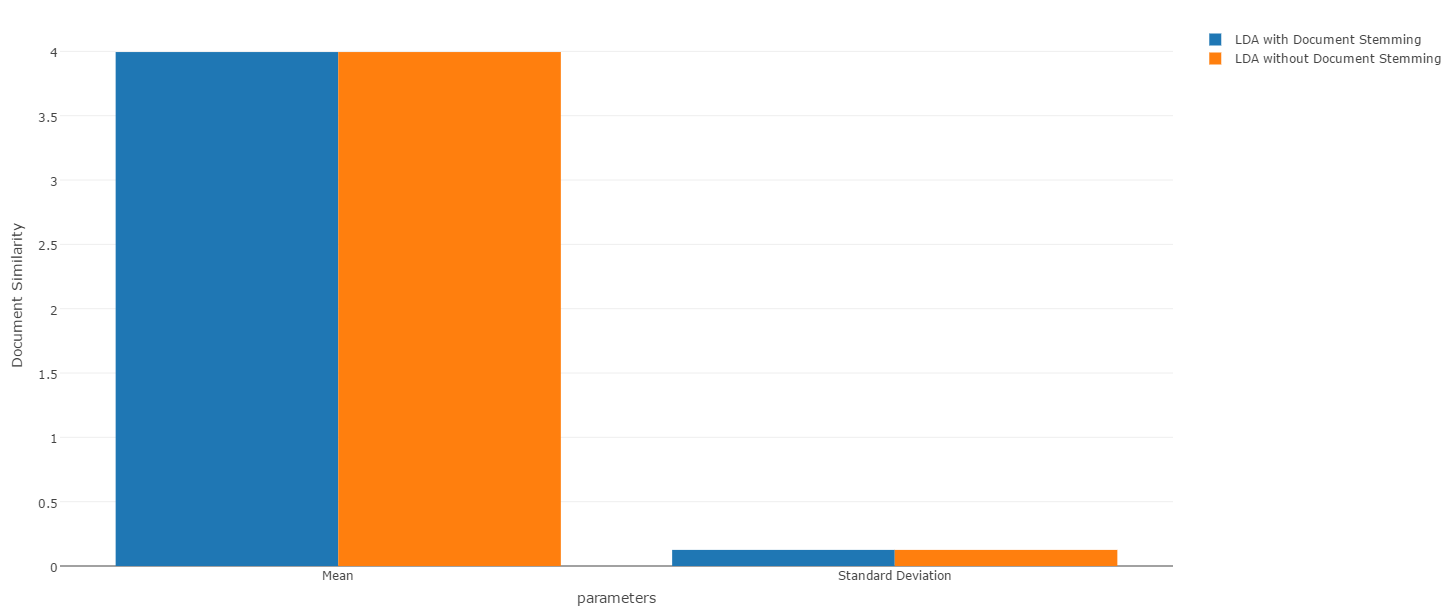
The LDA similarity is probably the worst choice to find the document similarity as the scores are almost similar and are in range of 3.999 in case of the number of topics are 8. The same case is in without stemming conditions, which is as shown in the below image.

*LDA similarity without Stemming*



As we can see there is not much difference in stemming and not stemming of the document. In both the cases, the similarity scores are almost similar and doesn’t show any significant results.

For emphasizing the point, below are the overall mean values and standard deviations for the LDA similarities. As we can see, the standard deviation is very low, which indicates the inefficiency of the method.

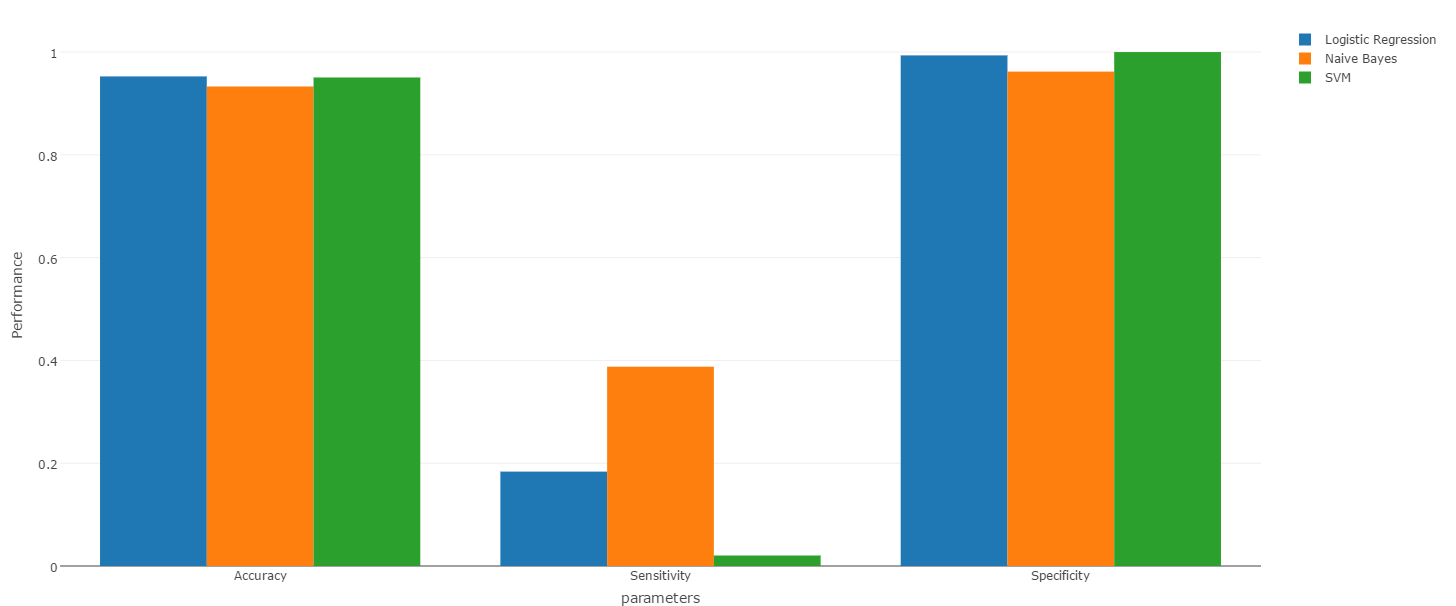


1. Predictive Model:

As already mentioned, the Quran verses were divided into 85% training and 15% testing data for finding a good classifier. Below is the overall performance of each methods in terms of accuracy, specificity, and sensitivity.

As shown in the figure, the accuracy and specificity of each classifier is about 95% and 97% respectively on an average. The deciding factor here is sensitivity, which is very low compared to the other two parameters. For the Naïve Bayes, the sensitivity is about 39%, while for the others it is below 20% and even 2% in case of SVMs. The primary reason for this low sensitivity is very low proportion of the data of violent verses, which is about 5% of the Quran verses. The performance can be lifted by feeding the classifier more inputs of the violent documents.

*Performance of classifiers*



**Conclusion and Future Work:**

It can be concluded that the documents used the verses of the Quran, but almost none of them were violent. Apart from that, one can identify by the predictive model, which verses are violent with the perspective of Quran.

The results found here are significant, though crude which suggest a great scope of improvements. It can be concluded that the cosine similarity is an efficient way to find the similarity between the documents, while the topic-modelling is not a good approach in such cases. In case of prediction, the Naïve Bayes can be considered as a most efficient classifier, which is performing better specifically in case of sensitivity.

In terms of the future work, some more expert help can be held for finding a more reliable score of Cosine Similarity, which would build more trust on the model. Some expert also suggests it to find similarity scores from some other Jihadi documents. Also, some more documents classified as ‘violent’ can be fed to the models to improve the sensitivity of it.