International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 10, Issue 1 (2024)
https://doi.org/10.31436/ijpcc.v10i1.321

Classifying Muslim Ideologies from Islamic Websites using Text Analysis Based on Naive Bayes and TF-IDF.

Salma Moustafa Sharey Moustafa, Akeem Olowolayemo*
Department of Computer Science, KICT, International Islamic University Malaysia.

*Corresponding author: akeem@iium.edu.my
(Received: 22nd April 2022; Accepted: 10th January 2023; Published on-line: 28th January 2024)

Abstract—Reliable digital Islamic information is one of the challenges faced by innocent Islamic information seekers such as young Muslims, new Muslims as well as others who desire to find authentic information about Islam, Prophet Muhammad (saw), and Muslims, in general. Several deviant ideologies abound, and they also present their information using the internet, sometimes involving digital deception. In the digital era, misleading Islamic information may affect people's beliefs, behaviours, and attitudes. Many websites are equally based on several schools of thought regarding Islamic practices which could be difficult for the new Muslims, and the young generation of Muslims to recognize what to follow among these different websites based on the information presented on the sites. Some other variants of practices are considered to be deviants by the mainstream Sunni scholars which may be misleading for innocent Islamic information seekers including non-Muslims. Consequently, the need to categorize different Islamic websites based on different schools and branches becomes imperative. This initial study focuses classification of Islamic websites utilising website categorization and text classification approach to their textual contents. The proposed technique classified 60 Islamic websites into two various categories Sunnah and Shia using TF-IDF for features extraction while using Multinomial Naive Bayes for classification. In addition, extracting the keywords for each of the two categories assisted in the classification process. The results show that Multinomial Naive Bayes was easily implemented and predicted the categories of Islamic websites with an accuracy of 0.89, precision 1.0, recall 0.80, as well as an F1 score of 0.89. The keywords that differentiate Sunnah's websites from Shia's websites were extracted. It was found that the best keywords that can be used in search engines to identify Sunni websites are Islam and Muslim, while Shia and Imam are the most prominent keywords that can be used to identify Shia's websites.

Keywords— Digital deception, Web Classifications, Text Classification, Content Filtering, Text Analysis, Naive Bayes, TF-IDF.

I. Introduction

Over the last decades, the world has witnessed a surge in the growth of the internet, with millions of web pages on every topic easily accessible through the web, making the web a huge repository of information. World Wide Web has made it essential for users to operate automated tools in finding the desired information resources. The World Wide Web collects unstructured, semi-organized, and structured text files, documents, photos, and other data types. With the rapid growth of the World Wide Web (WWW), there is a growing need to provide web users with automated web page categorization to facilitate the indexing, browsing, and retrieval of pages. To achieve the web's full potentials as an information resource, the vast amount of contents available on the internet have to be well described and organized. Such assistance is useful in the organization or development of the catalogs organizing web documents for large numbers of the information returned by keyword search engines. However, this is difficult to meet without automated web-page classification techniques due to the labor-intensive nature of human editing. That is why automation of web page classification is useful.

Web page classification helps in focused crawling, develops and expands web directories, helps in analyzing specific web link topics, analysis of web content structure, increases the efficiency of web search (e.g. category view, ranking view), web content filtering, web browsing assistance, and more [1]. Many types of features are extracted from a website to conduct a webpage classification. For example, features can be extracted from texts or images or URLs, based on the intended approaches, namely; content-based classification, functional-based classification, and sentiment-based classification. With the ever-increasing nature of textual material stored electronically on the Internet, automated text classification is now an essential application of machine learning. The more common approach used was using the text content of the web page to classify its topic.

The emergence of computer technology, internet and digital information has a big impact on the Muslims just like the impacts on everyone on the planet. In recent years, the Internet has become an important medium for Muslims to pursue and share information about Islam. The need for Islamic information and awareness online has become common among Muslim and non-Muslim users interested in Islam. Fake or misleading information regarding religious issues can influence people's beliefs, attitudes, intentions, motivations, or behaviour. It can distort the true Islamic teachings, especially for new Muslims and the younger generation [2]. Islamic websites are very useful for new Muslims and non-Muslims who want to understand Islam in a simple and systematic way. But Muslims have different divisions, some of them are considered to be deviant ideologies that may influence the internet users' beliefs, behaviours, and practices. Consequently, it is necessary to categorize different website based on schools and branches of Islam, especially to identify acceptable variations as well as deviant thoughts, opinions or practices.

This study is motivated from the forgoing and aims to introduce a method to classify Islamic websites from two different Islamic branches, classified broadly into Sunni and Shia. The techniques employed precisely is Naïve Bayes algorithm in conjunction with Term Frequency-Inverse Document Frequency (TF-IDF) with keywords extracted from each branch to facilitate effective categorization of the websites and assist in getting the right Islamic content through web search engines.

Deviant Islamic teachings are teachings and practices that contradict Al-Qur'an and the Sunnah as well as contradict the mainstream ideology or opinions of Sunni muslim scholars [3]. It could be difficult to define truth in the context of Islamic websites because there are many different schools of thought regarding Islamic practices [4]. As a result, the teachings of Islam for new Muslims and the younger generation can be altered [2]. The main branches within Islam are Sunnah and Shia. Classifying these two broad category is paramount especially based on general views of scholars. For instance, mainstream scholars such as Malaysian Islamic scholars consider Shia schools to be deviant and positioned Shia, liberalism, and pluralism among deviant ideologies according to Islamic teachings [5]. Consequently, sought to prohibit Shia propaganda and indoctrinations in journalism and mass media [6] .

Moreover, many parents are afraid that their children may unknowingly be indoctrinated by deviant websites based on distorted or deviant information. Hence, they would prefer to filter out these contents to prevent their children from viewing or accessing these websites [4], [7] . Hence, the prospect implication of this study is to propose a system to help Muslim parents to set up control of the digital Islamic content for their children to help them maintain mainstream Islamic contents based on their branch of Islam as well as their preferred school of thought and to prohibit them from viewing deviant Islamic websites. This should also be useful for new Muslims, the young Muslim generation, and non-Muslims to distinguish among the different ideologies.

This study focused only on categorizing the websites based on their content according to only the textual contents. Sunnah and Shia websites were categorized in this study to differentiate between them. Other divisions were not considered in this study.

The remaining parts of this study is organized as follows. The next section presents a review of related work, which was followed by the section on the methods and the steps undertaking for the classification. Subsequently, the results of the experiment are discussed and the last section presents the conclusion of the study, some limitations, and highlights directions for future work.

II. Related Work

Many approaches for website categorization abound, specifically subject-based classification which categorizes the content or subject of the webpage, functional-based classification which aims to discern the purpose of the website, and sentiment-based classification that focuses on the opinion expressed on the web page. Each approach can be done according to visual content or textual content or combining visual and textual contents, or utilizing the URLs of the websites.

Website classification or website categorisation is defined as the classification of websites into appropriate categories based on their contents and intents. It assigns a web page to one or more predefined labels in a category that plays an important role in oriented crawling, assisted web directory development, topic-specific web link analysis, context-based advertisements, and web-structure analyses. Website categorisation can be achieved on a wide scale and wide variety of usage using specialized tools or by manually accessing big data of online domains.

Website categorization is useful to detect and prevent insider risks. For instance, it is customary for companies to block entertainment sites, social media, block special sites for security reasons. Those sites often involve malicious material, adult content, and phishing sites. This technique can also be applicable to parental controls. Parents can choose which categories of websites they want to prohibit their children from viewing. They may include sites with pornographic content, vulgar language, those selling illegal software, guns, and drugs, or websites promoting abuse and cruelty. Parents may also restrict their children's access to social networks, emails, online shops, e-payment systems, and other services.

https://doi.org/10.31436/ijpcc.v10i1.321

Subject-based classification is one of these approaches in which Web pages are categorized based on their contents or subject matters, e.g. "Religion," "Education," "Entertainment," "Sports," "News," "Blogs" or "Science." It may be used to build topic hierarchies of web pages and, subsequently, to conduct context-based searches for web pages relating to specific subjects. Another classification approach is the functional-based classification focusing on the intent of the website. e.g. "Personal homepage", "Course page", "Admission page", and so on. Sentiment-based classification focuses on the opinions expressed on the web page, that is, the author's attitude to a specific topic. Several methods are employed for websites classification. Text-based classification method is the process of assigning tags or categories to a website according to its textual content. Another classification method is the image-based classification that classifies websites using visual content. The method of combining visual and textual characteristics for website classification is the combined text and image-based classification. Link-based classification method is based on the information from the hyperlinks of a webpage. Text classification (text categorisation) is a process by which tags or categories are assigned to text according to its content. It is an essential task in the processing of natural language NLP with wide applications including sentiment analysis, subject labelling, spam detection and purpose detection. The text classifier can be used to organize, structure and categorize almost any type of text from materials or files throughout the web. Many phishing websites are designed to be like the legitimate websites to facilitate deceiving users and gain access to their sensitive and valuable data, hence, crucial to identify these phishing websites automatically.

Previous studies such as in [8] used different machine learning algorithms, namely; Naive Bayes, SVM, Decision tree and the neural network using MATLAB script to compare the performance of classifying 1,353 URLs to detect phishing websites. They extracted 9 features from each URL. The results show that the accuracy of the decision tree classifier is poor because of the overfitting of the classifier so they had to prune the classifier to improve the accuracy to 90.39%. The best classification performance was achieved from the pruned decision tree while neural network classifier achieved the lowest accuracy with average performance of 84.87% [8].

Another study in [9] wanted to detect spam comments on YouTube videos by implementing different supervised classification algorithms, single classifier algorithms such as Naive Bayes, K Nearest Neighbour (KNN), and Support Vector Machine as well as ensemble classifier (Bagging). The experiment is executed using Weka machine learning tool. The classification is binary, specifically classifying the comments into two categories (spam or not spam). Five datasets were collected from UCI data repository and preprocessed the extracted features using Bag of Words and TF-IDF to improve the classification effectiveness. The performance result for most of the algorithms is above 80%. The highest accuracies were achieved from ensemble classifier (Bagging) and Naïve Bayes classifier. The accuracy of 1-neighbor KNN was higher than that of the 3-neighbors KNN [9].

Furthermore, the researchers in [10] implemented Naïve Bayes classifier to detect fake news. They collected 2282 Facebook posts represented by news articles, filtered them into 1771 news articles based on the contents and relevant label. The dataset was shuffled randomly and split into training, test and validation sets. Training categorized the dataset into three classes, namely; true news, false news and fake news. Fake news accounted for about 4.9% of the whole dataset. The classification accuracy was 75.40% with an average precision of 0.71 and a recall of 0.13 [10].

One type of web threat that deceives users to steal their private and sensitive data and also may yield financial losses is phishing websites which are designed to look like legitimate webs. This kind of threat became a serious issue that attracts many researchers to find a method to detect it automatically. Subasi et al.[11], implemented different machine learning tools namely Random Forest, Support Vector Machine SVM, Artificial Neural Networks ANN, k-Nearest Neighbour KNN, C4.5 Decision Tree, and Rotation Forest RF by using the WEKA2 tool to detect the phishing websites. They collected datasets from the UCI machine learning repository. The results showed that all machine learning techniques performed well in detecting phishing websites with accuracy from highest to lowest 97.36%, 97.18 %, 97.17%, 96.91%, 96.79% and 95.88% for Random forests, k-NN, SVM, ANN, RF and C4.5 respectively. For the whole algorithms, F-measure is equally coincident with the accuracy performance, specifically; 0.974, 0.972, 0.972, 0.969, 0.968, 0.959 respectively based on the aforementioned accuracies. The highest, fastest, and most efficient performance was the random forest algorithm.

Abdallah & Iglesia [12] implemented a Naïve Bayes classifier for classifying websites based on their URLs utilizing the DMOZ dataset containing 15 categories of 1562808 URLs. For feature extraction, they proposed an n-gram character-based language model. The results of their classification method showed an F-measure of 82.72%.

The study in [13] was based on research from the previous paper [12] that tried to improve its classifying performance. The study in [13] implemented Multinomial Naïve Bayes. Preprocessed the datasets by removing all kinds of punctuations and stop words to get rid of unnecessary words and decrease noise in the dataset. The feature

https://doi.org/10.31436/ijpcc.v10i1.321

extraction technique was improved by combined n-gram with the TF-IDF technique. They used unigram (n=1) and bigram (n=2) to extract some words from URLs and counting their occurrence number then building the vocabulary, followed by TF-IDF for features extraction to determine the most relevant words in each category. Datasets were split into training and test set with 1532808 URLs and 30000 URLs respectively. To measure the performance of the classifier, average precision, recall, and F1-score were examined, giving average results of 91.30%, 88.77%, and 87.63% respectively [13].

A hard categorisation approach has been adopted to classify the website into more than one category [14]. It was usually possible to know the purpose of any website using the information contained on the homepage. Hence, the authors used a dataset of 319 websites homepage and classified the websites into seven categories. They utilized the main features of the homepage like the meta tags, title tag, heading tags, hyperlinks, the website's content and the domain name of the website as features for the classification. The approach used was a keyword-based algorithm that analysed the root of each keyword. This approach enabled the researchers to identify which keywords were relevant for which category. They assigned a score for each category to classify each website, placing the websites into categories based on the highest score. Precision, recall and the F-measure were measured to assess the efficiency of the techniques, showing an average precision of 95.2%, an average recall of 94.6% and an average F-measure of 94.7% [14].

In [15], the researchers attempted Naïve Bayes approach to classify websites using their URLs. The system was designed to classify the URL in one of the 9 categories, i.e. Arts, Business, Computers, Games, Health, Home, News, Recreation and Reference. A total of 8,55,939 URLs of 9 categories have been extracted from the Open Directory Study (ODP) dataset. 90% of the training set was used and tested to form a dictionary in each category, a total of 9 dictionaries were used to train the Bayes classifier. The URLs from each category were parsed into tokens and each token was checked with all the dictionaries in the training set. Common words or strings such as "www", "http", "html" etc. were removed from the list of tokens to construct the dictionary based only on the unique tokens. The probability that the given URL belongs to a particular category were computed by applying the Bayes theorem. In each dictionary, the occurrence of these tokens was verified and the number of equal matches and partial matches counted. Precision, Recall and F-Measure values have been evaluated to determine the performance of the classifiers. The average value of Precision, Recall and F-measure of 0.7, 0.88, and 0.76 respectively were achieved [15].

This related work section began with the description of different concepts related to webpage categorisation and text classification following by reviewing relevant literature and past studies that were conducted to generate understanding in the area related to text classification. The following section presents the methods for the study.

III. Methods

In this study, an experiment to classify Islamic websites into two categories, Sunnah and Shia, based on their textual contents was conducted. Multinomial Naive Bayes classifier will be used to conduct the classification experiment. To improve the classifier performance, TF-IDF technique will be used to extract the features from text data. The data will be split into 70% and 30% for training and testing data respectively. Then the classifier will be applied on the training set followed by the test set. The following figure will explain the steps of conducting the classification experiment, also the steps will be explained in detail in the subsequent sections.

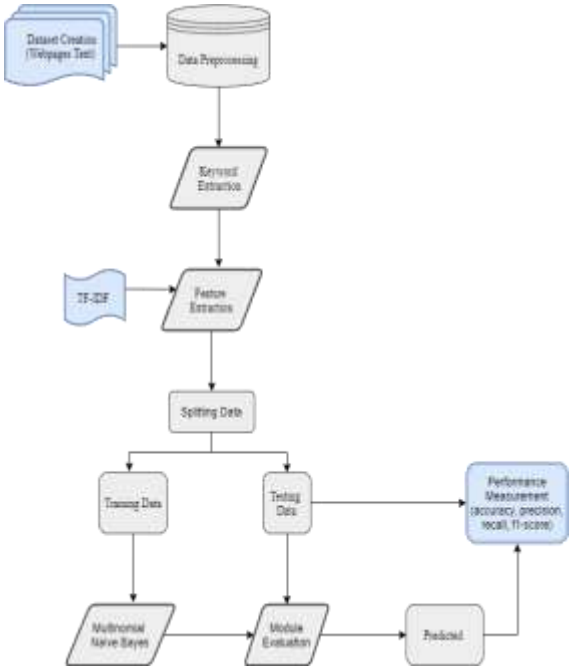

Figure 1. Webpages Classification Flowchart

A) *Dataset Collection*

Table I
URLs for both categories

Shia	Sunnah
https://www.sistani.org/english/	https://www.arabmediasociety.com/
http://thetrueshia.com/	https://islamqa.info/en
http://www.shia.org/	https://www.alim.org/

https://doi.org/10.31436/ijpcc.v10i1.321

https://www.al-khoei.org/	https://en.islamway.net/
https://shiacourse.urd.ac.ir/	https://www.islamweb.net/en/
https://shia-online.com/	https://www.islamreligion.com/
https://majlis.org.uk/	https://www.islamicity.org/
http://www.shiasearch.com/en	https://www.dar-alifta.org/
http://shiarightswatch.org/	Foreign/default.aspx
https://www.imams.us/	https://islamicfiqh.net/en
https://al-rida.net/	https://www.iifa-aifi.org/en
https://shiamuslimcouncil.org/	https://sunnah.com/
http://ishia.org/	https://islamhouse.com/en/source/9203/
https://imam-us.org/	https://www.nabulsi.com/web/en
https://al-mostabserin.com/english/	https://free-islamic-course.org/
https://www.almahdi.edu/	https://www.muis.gov.sg/
https://www.alzahracentre.org/	https://muslim.sg/
http://shiastudies.org/home	https://islamstory.com/en
http://www.jammat.org/	https://alfarooqcentre.com/en/
https://www.orlandocenter.org/	https://islamfaith.com/
https://www.sijny.org/	https://www.quranicthought.com/
https://www.ihcc.org.au/	https://islamireland.ie/
https://www.orlandocenter.org/	https://sunnahonline.com/
http://www.dar-ul-hikmah.org/	https://www.fatwa-online.com/
https://shia-calendar.com/	https://www.whyislam.org/
https://nazmia.org/index.htm	https://muslimheritage.com/
https://www.tnice.co.uk/	http://www.islambasics.com/
https://masom.com/our-mission	http://www.islamtomorrow.com/
https://www.shiamatch.com/	https://www.discoverislam.com/
https://www.aljaafaria.com.au/	https://www.islaam.net/
	https://understandquran.com/

The data for the experiment have been collected from various Islamic websites, some of them are Sunni's websites and the remaining are Shia's websites, as presented in Table I. These websites are found by following the instruction in the section title "Representing Shia and Sunnah online", crawling and extracting from several Islamic websites using the keywords such as Sunnah, Sunni, hadith, fiqh, Islam, Muhammad, and Quran Shia, Imam, Imamah, Twelver, Shiite etc. The crawled websites were then categorised and labelled Sunni's websites, and Shia's websites with the help of an expert familiar with the contents on the websites. The text data extracted were from the first two pages from each website, namely; the homepage and the next. A total of 60 websites were crawled.

The study was designed to classify the websites based on their textual contents into one of 2 categories, "Sunnah" and "Shia". Textual dataset obtained from these webpages were cleaned to get rid of unnecessary parts of data and remove noise by changing all characters into lowercase, removing punctuations marks, removing numbers and stop words as well as converted into tokens, and subsequently features were extracted to form the dataset partitioned into training and testing sets for the classification process.

As it is usually possible to know the purpose of any website using the information contained on the homepage, so, datasets that exist in homepages were extracted for training the classifier. Information contained in other pages from each of the websites were added to improve the performance of the classification and subsequently, to attempt to classify the websites into two categories.

B) Keywords Extraction

Extracting keywords of any category is important in classifying websites. In this study, the keywords are extracted automatically from the selected websites that have already been categorized into Sunnah and Shia and the most occurring words are counted in each category. The pool of the extracted keywords is presented in Table II.

TABLE II
SAMPLE OF THE PREPROCESSING EXTRACTED KEYWORDS

[aal', 'ability', 'able', 'absolute', 'absolute', 'abulqassim', 'accept', 'accept', 'accomplish', 'according', 'according', 'according', 'according', 'accordingly', 'account', 'achieved', 'achievements', 'acquaintance', 'acquainted', 'acquainted', 'activities', 'ad', 'ad', 'addition', 'admitted', 'advised', 'affect', 'age', 'age', 'agha', 'agha', 'ago', 'ago', 'agree', 'agreed', 'ah', 'ah', 'ah', 'ah', 'ah', 'ah', 'ah', 'ah', 'ah', 'ah', 'ah', 'ahlul', 'ahmad', 'ahmad', 'ahvaz', 'aim', 'alalaam', 'alansari', 'alawwal', 'aletikaaf', 'alfadhairi', 'alfadhairis', 'alfyyah', 'alhakim', 'alhakims', 'alhussani', 'alhussein', 'alhusseini', 'alhusseini', 'alhusseini', 'alhusseini', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'ali', 'alim', 'alim', 'alims', 'alive', 'alive', 'alkhadra', 'alkhui', 'alkhui', 'alkhuis', 'alkhuis', 'alkifayas', 'allah', 'allah', 'allah', 'allah', 'allah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'allamah', 'almakasib', 'almohri', 'alnizam', 'aloof', 'alsadr', 'alsaduq', 'alsawm', ……..

C) Feature Extraction

Feature extraction is critical in text classification because it has a direct impact on classification accuracy. Pre-processing data cannot be classified by the classifier until transforming it into a set of features. In this study, the classification is binary as there are just two categories (Sunnah, Shia) in which the websites are to be classified, so the Sunnah category was denoted by "1" and Shia category by "0" to facilitate numeric identification by the classifier. Then, the pre-processed data was transformed into features using the technique of term frequency-inverse document frequency (TF-IDF) that was built by utilizing the sklearn library. It was used to evaluate the words used in a document and compares their usage to the words used in the whole documents. A score was assigned to emphasises each word's importance throughout the document.

https://doi.org/10.31436/ijpcc.v10i1.321

D) Algorithm Selection

In-text classification, the selection of the suitable machine learning algorithm based on various factors including; training data size, speed of training, and output accuracy. In this study, the Multinomial Naïve Bayes algorithm was used because it is a predicting algorithm that predicts the category of each website. It can perform well with small sample size. It also produces good accuracy relative to the number of trained data. It also classifying data rapidly, accurately and can be implemented easily [16]. The multinomial model was designed to categorize documents based on the word frequency in it, but some words may not add meaning to the document but have high term frequency [17] . To overcome this problem in the model, TF-IDF was used as a feature extraction technique in this study to access the most relevant word as well as remove unnecessary words in each document.

E) Split Data into Train and Test Set

Multinomial Naïve Bayes is a supervised learning algorithm in which the machine gets its knowledge from the labelled training data and then use this knowledge to predict the output of the other data, so the data have to be split into training data and testing data. From the sklearn library, train_test_split () is used to split data into training data and testing data with a percentage of 70% and 30% respectively.

After splitting, the model was applied using sklearn Naïve Bayes library on the training set and then a prediction on the testing set was performed.

This section analyzed the appropriate methods that were applied to classify the crawled Islamic websites into two classes, namely Sunni and Shia. The steps involved in the experiment that were conducted to achieve the purpose of this research were analyzed in detail.

IV. Evaluation

This section presents the experimental results from classifying the 60 Islamic websites into two categories Shia and Sunnah, which is the main objective of this study.

A) Keywords of each category

Keyword extraction is one of the crucial ways for analyzing different categories' data. The keywords highlight the main ideas of each website and help people to access the right website through keyword search engines.

After dataset preprocessing, the keywords were extracted. Table III shows the ranked keywords for each category and the number of occurrences of each keyword. The two most important keywords that can be used to search for Sunni websites are Islam and Muslim while for Shia's websites, the most prominent keywords are Shia and Imam.

Table III
Keywords of Each Category and Their Number of Occurrence

Sunnah		**Shia**	
Keyword	frequency	Keyword	frequency
Islam	679	shia	420
muslim	605	imam	404
islamic	519	prayer	319
quran	350	islamic	263
god	258	surah	200
allah	246	recite	192
prophet	226	holy	147
muhammad	197	islam	142
prayer	135	shaykh	135
mosque	98	religious	132
hajj	97	allah	122
hadith	87	ali	177

B) Performance Measure of the Classifier

As shown in Table IV, the highest performance is achieved when splitting data with 70 % and 30% for the training and test sets respectively, in the evaluation stage of the Multinomial Naïve Bayes model, specifically, splitting the dataset into 42 training websites and 18 test websites.

Table IV
Performance Measurement of Different Splitting Values

Splitting Value (Training, Testing)	**Accuracy**	**Average Precision**	**Average Recall**
90% , 10%	0.83	1.00	0.75
80% , 20%	0.75	1.00	0.63
70% , 30%	0.89	1.00	0.80
65% , 35%	0.81	0.89	0.78
60% , 40%	0.63	1.00	0.34

C) Testing Classifier

To predict the category of the Islamic websites, the Multinomial Naïve Bayes algorithm was trained by using the training set and then using the output to classify the test set and predict its categories. The predicted categories of the test websites are presented in Table V.

Table V
Performance of Test Set for the Classifier

ytest	1	0	1	1	0	1	1	0	1	1	0	1	1	0	0	0	0	1
predicted	1	0	0	1	0	1	1	0	1	1	0	0	1	0	0	0	0	1

Sunnah website is represented by (1) and the Shia website is represented by (0), the above table shows that there are (2) incorrect prediction categories and (16) correct prediction categories. The number of true Sunnah predictions is 8 and the number of true Shia predictions is 8.

The performance of the classifier was carried out based on the accuracy, precision, recall and f-measure using the actual and predicted values to gauge the performance of the classifier. The best performance are based on 70%-30% split.

https://doi.org/10.31436/ijpcc.v10i1.321

The average recall was 0.80 which indicates that the classifier is good in classifying data into any specific category. The result of measuring the precision is 1.00 which demonstrates that the performance of predicting any category is quite high. The average accuracy of the classifier is 0.89 which refer to how suitable is this classifier for predicting the category of any Islamic website. Precision, Recall and F-measure values for each category were measured and their results are shown in Table VI.

TABLE VI
PERFORMANCE MEASURES OF 60 ISLAMIC WEBSITES

	Sunnah	**Shia**
Precision	1.00	0.80
Recall	0.80	1.00
F1-score	0.89	0.89

The speed of the classifier was measured by measuring the training and testing time. The training time for the Naïve Bayes classifier is 0.000000 sec and to predict the category of the test data the testing time was 0.000000 sec. These results suggest that Multinomial Naïve Bayes speed is efficient for classifying websites using textual data.

V. CONCLUSIONS AND FUTURE WORK

The digital age has witnessed an emergence of Islamic learning practices on the internet. The internet or specifically, the web, is a vital medium for conveying information and knowledge about Islam. This knowledge ranging from the fatwa, aqeedah, fiqh, tajweed, hadith, tafseer, beliefs, practices, Islamic history and so on. It can be presented through media such as web text, audio lectures, and videos. Although the advantages of digital Islamic learning are quite obvious, there are downsides as well. Sometimes, there is deviation or distortion of the true Islamic teaching in the digital Islamic contents as users can use the internet to publish inauthentic, fabricated or misleading information that is not based on the Quran and reliable Sunnah. As a result, there is the need to classify the Islamic websites to facilitate identifying and classifying the deviant websites as well as to simplify the access to the reliable, mainstream online Islamic contents. This cannot be done effectively without automated classification techniques.

In this study, 60 Islamic websites were selected from different two categories, namely; Sunnah and Shia to classify their text content automatically. Keywords are extracted from these websites which will facilitate classifying websites into categories that facilitate access to the right Islamic content and identify the deviant Islamic website. Moreover, supervised machine learning algorithms, specifically, Multinomial Naïve Bayes algorithm which uses pre-classified data for training were employed to train the classifier and to determine the performance of the classifier using the results on the testing data to predict their categories. This algorithm was based on the feature extraction technique TF-IDF to improve the performance of the categorisation.

The results showed that Multinomial Naïve Bayes was implemented easily and predicted the category of the Islamic websites rapidly and accurately. The best result was achieved with a splitting value of 70% - 30% for training and test sets data. The testing data had an accuracy of 89%, an average precision of 1.0, as well as an average recall of 0.80. The keywords that distinguished between Sunnah's websites and Shia's websites were extracted. The top keywords that can be used to search for Sunni websites are Islam and Muslim, while Shia and Imam are the top ranked keywords for Shia's websites.

Hence, Multinomial Naïve Bayes can be used in the classification of texts from Islamic website as the accuracy of the classification of the trained classifier is reasonable for the test collection.

A) Limitation

The first limitation is related to the keyword extraction process. The disadvantage of this method is that all frequent words are considered equally important, although there are some words such as part, books, video, read, contacts and article which can appear frequently and commonly found in all categories. They, however, have little importance and have no effect on the kind of classification of the categories. These words should not be considered as category keywords. Consequently, attempts were made to exclude them from the list of keywords.

The data size was 60 websites where 42 websites were utilised for training the classifier with the performance accuracy of the classification reaching 89%. It is believed that increased dataset would likely improve the results, since the size of the training dataset can affect the performance accuracy of the classifier.

B) Further Work

Due to the mentioned limitations of this study, future studies should use TF-IDF to extract the important and most significant keywords from the document because it accesses the relevance of a word to any document. By getting the highest TF-IDF score for each category, getting the most accurate keywords is more achievable. Moreover, the size of datasets should be increased to improve the performance of the classifier. Also, sentiment analysis may equally be introduced to categorize websites, where its performance can be compared to the present results. Classifying all Islamic ideologies should be conducted, for example; Sunni websites and non-Sunni websites as well as categorizing the websites according to core specialties such as kids, adults, fiqh, aqidah, civilization, beginners, intermediate and experts which should also include an

approach to rate the quality and reliability of the information on the websites.

ACKNOWLEDGEMENT

We would like to express our gratitude to those that participate in this research

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

[1] L. Safae, B. El Habib, and T. Abderrahim, "A Review of Machine Learning Algorithms for Web Page Classification," in *5th International Congress on Information Science and Technology (CiSt)*, 2018, pp. 220–226.

[2] F. A. Mohamed, M. S. Abdul Aziz, M. Mahmud, and Z. Zulkifli, "Identifying Cues to Deception in Islamic Websites Text-Based Content and Design," in *International Conference on Information and Communication Technology for the Muslim World*, 2018, pp. 285–289.

[3] A. Idris and O. Kurtbag, "A Comparative Study of Government Policy in Dealing with Deviant Teachings in Islam: The Case of Malaysia and Turkey," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. Vol. 9, N, 2019.

[4] M. Mahmud and A. Abubakar, "Investigating the Act of Deception in Online-Islamic Content," in *2014 3rd International Conference on User Science and Engineering (i-USEr)*, 2014, pp. 1–6.

[5] N. Saat, "Johor and Traditionalist Islam: What This Means for Malaysia," in *Southeast Asian Affairs; Singapore*, 2018, pp. 186–200.

[6] V. Kirichenko, "The Shia Community IN Malaysia," *Russ. Moslem World*, vol. N 1 (307), pp. 80–91, 2020.

[7] L. Roberts and R. Samani, "Digital Deception: The Online Behavior of Teens," pp. 1–8, 2013.

[8] A. D. Kulkarni and L. L. Brown III, "Phishing Websites Detection Using Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. Vol. 10, 2019.

[9] Z. Zaman and S. Sharmin, "Spam Detection in Social Media Employing Machine Learning Tool for Text Mining," in *13th International Conference on Signal-Image Technology and Internet-Based Sys*, 2017, pp. 137–142.

[10] M. Granik and V. Mesyura, "Fake News Detection Using Naive Bayes Classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 2017, pp. 900–903.

[11] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, "Intelligent Phishing Website Detection Using Random Forest Classifier," in *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2017, pp. 1–5.

[12] T. A. Abdallah and B. de La Iglesia, "URL-Based Web Page Classification: With n-Gram Language Models," in *In International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, 2014, pp. 19–33.

[13] A. Shawon, S. T. Zuhori, F. Mahmud, and M. J.-U. Rahman, "Website Classification Using Word Based Multiple N-Gram Models And Random Search Oriented Feature Parameters," in *2018 21st International Conference of Computer and Information Technology*, 2018, pp. 1–6.

[14] S. Pudaruth, Y. Ankiah, and K. Sembhoo, "Using a Thesaurus-Based Approach for the Categorisation of Web Sites," 2014.

[15] R. Rajalakshmi and C. Aravindan, "Naive Bayes Approach for Website Classification," in *International Conference on Advances in Information Technology and Mobile Communication*, 2011, pp. 323–326.

[16] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory," pp. 1–20, 2014.

[17] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison Between Multinomial and Bernoulli Naïve Bayes for Text Classification," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, 2019, pp. 593–596.