Arabic Text-Based Chat Topic Classification Using Discrete Wavelet Transform

Arwa Diwali[1], Dr. Mahmod Kamel[2] and Dr. Mohmmed Dahab [3]

**[1] Information Systems Department, King Abdulaziz University, Jeddah,
Saudi Arabia**

**[2] Information Systems Department, King Abdulaziz University, Jeddah,
Saudi Arabia**

**[3] Computer Science Department, King Abdulaziz University, Jeddah,
Saudi Arabia**

Abstract

Research studies on topic classification of chat conversations use the Vector Space Model (VSM) extensively. The VSM represents documents and queries as vectors of features. These features are the terms that occur within the collection. The VSM's limitation is that it does not consider the proximity of the terms in the document. The proximity information of the terms is an important factor that determines their position in the document.

Another model used in information retrieval systems is the Spectral-Based Information Retrieval Method (SBIRM), which employs the Discrete Wavelet Transform (DWT) to rank documents according to document scores. This method's advantage is that it not only counts the frequency of the terms used in the document, but also considers their proximity by comparing the query terms in their spectral domain rather than their spatial domain.

Based on the foregoing considerations, the objective of this research is to build a framework for Arabic Chat Classification (ACC) that can help detect illegal topics in Arabic text-based chat conversations. This framework is a combination of Information Retrieval (IR) and Machine Learning (ML) methods. The ACC implements two methods: first, the SBIRM using the DWT and second, the Naïve Bayes method.

Two experiments were conducted to test the ACC framework, one for root-based and the other for stem-based chat conversations. The results showed that the former outperformed the latter in terms of accuracy, precision, and F-measure. The recall results were the same for both experiments.

Keywords: *Chat Classification, Discrete Wavelet Transform, Naïve Bayes, Term Signal, Spectral Based Retrieval Method.*

1. Introduction

The Internet is a powerful tool of communication and collaboration among people around the world. With the increased availability of Internet access through computers, smart phones, and handheld devices, chat conversations are an easy way of meeting new people online and interacting with friends and relatives; unfortunately, online chats have also become a target of illegal activities. Chat conversations, which take place in chat media, can contain important information regarding participants, such as their behaviors, intentions, habits, and tendencies. Hence, it is necessary to analyze chat conversations and to understand what people are looking for and what knowledge they exchange with each other for economic benefit, legal purposes etc.

Research studies on the topic of classifying chat conversations [1], [2], [3], [4], [5], [6] extensively use the VSM with different weighting schemes. One of the techniques most frequently used in information retrieval systems, the VSM represents documents and queries as vectors of features [7]. These features are the terms that occur within the data set, with the individual value of each feature representing the occurrence or non-occurrence of the term within the document that it represents. If there are N terms in a document data set, then each feature vector correspondingly contains N dimensions.

In its simplest form, the feature value may use a binary value to point out the existence of a term. An improved model may include the frequency of a term, based on the assumption that the more often a term is used, the greater its importance to the document. This often has the unfortunate side effect of lending too much weight to common terms that may occur with a high degree of frequency throughout the entire data set, so schemes such as Term Frequency X Inverse Document Frequency (TF X IDF) are used to discount these high-frequency terms.

A major limitation of the VSM is that it does not take into consideration the proximity of the terms in the document. The proximity information of the terms is an important factor that determines their position in the document. Based on the basic linguistic assumption, the proximity of the terms in a document implies a relationship between those terms. Given sentences that contain a single idea, or a cluster of related ideas within neighboring sentences, there is a probability within the document structure that terms used together are related. On the other hand, when two terms are on opposite ends, the probability of a relationship between the terms is relatively weak. The proximity of the terms is as significant a factor as is their frequency, and which must not be ignored in the topic classification of chat conversations. In contrast, the Spectral-Based Information Retrieval Method (SBIRM) is another model used in information retrieval systems to rank the documents according to document scores. Introduced by[8], this model combines the frequency and proximity of the query terms. The SBIRM is able to overcome the VSM limitation by comparing the query terms in their spectral domain rather than their spatial domain. Previous research has shown that SBIRM increases the average precision when compared to corresponding the VSM [9]. Furthermore the chat conversation is compact and the proximity factor is must take in to consideration. To analyze the relative positions in SBIRM, the vectors are mapped into the frequency domain. The term position is treated as the position in time. Performing a mathematical transform such as a Fourier Transform [10], Discrete Cosine Transform (DCT) [11], or Discrete Wavelet Transform (DWT) [9], allows us to observe the term spectrum in relation to a certain document. The DWT's advantage is its ability to break a signal into wavelets of different scales and positions so that it can analyze the patterns of the terms in the document at various resolutions (whole, halves, quarters, or eighths).

While most studies on Chat Mining (CM) have been conducted in English and European languages, this research focuses on the Arabic text-based chat conversations. Due to the huge amount of Arabic chat conversations that appear on the Internet, Arabic CM becomes necessary to make sense of all this information and to perform knowledge discovery from collections of unstructured Arabic text-based chat conversations.

Our research aims to design a prototype of an Arabic Chat Classification (ACC) framework that can assist in detecting illegal topics in Arabic text-based chat conversations. This framework involves a hybrid approach, which combines Information Retrieval (IR) and Machine Learning (ML) methods. The ACC implements two methods: first, the SBIRM using the DWT [9] to overcome the VSM limitation by comparing the keyword terms in their spectral domain rather than their spatial domain, and second, the Naïve Bayes (NB) method to classify the topics of Arabic text-based chat conversations. With this merging, we can benefit from the document scores of the SBIRM to improve the performance of the classifier and to increase the precision of the predicted class. In addition, the research objective is to investigate the impact of text preprocessing techniques

The rest of this paper is organized as follows. The literature review is given in Section 2, and ACC framework implementation is presented in Section 3. The experiments and results are discussed in Section 4. The conclusion and future work are given in Section 5.

2. Literature Review

Various studies were conducted in chat mining topic detection using VSM. The study of [2] assisted crime detection and prevention by automatically creating concept-based profiles that summarized chat session data to detect topics. The researchers created ChatTrack, a system that can generate a profile of the topics being discussed in a particular chat room or by a particular individual. For each category, the classifier creates a vector of representative keywords and their weights based upon (TF X IDF) formula. The work by [5] proposed a system called IMAnalysis, which supports intelligent chat message analysis using text mining techniques. They used supervise topic detection approaches like NB, associative classification, and Support Vector Machine (SVM). The IMAnalysis system provides the following three functions: chat message retrieval, social network analysis, and topic analysis. Chat message retrieval combines general browsing and the retrieval of chat sessions. It also shows statistical information about chat activities, such as the average number of messages and the average number of words in the message. Social network analysis discovers the social interactions of Instant Messaging (IM) users and their contacts. Chat topic analysis automatically detects the topics with which IM users are involved. The topics are limited to five categories: sports, games, entertainment, travel, and pornography. The dataset for the five categories was collected from several chat websites.

IMAnalysis is capable of classifying chat sessions with multiple class labels using the Indicative Terms Dictionary. Based on researchers' observations of chat conversations, the Indicative Terms Dictionary contains a set of words called indicative terms or topic keywords that characterize a specific topic. The results show that the SVM classifier outperforms the other two classifiers in precision, F-measure, and accuracy.

The researcher in [6] conducted a study to determine the topics in Turkish text-based chat conversations. The supervised learning methods ware used in this study. The data set was represented using the (TF X IDF) weighting scheme.

A common formula for (TF X IDF) is given as Eq. (1).

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{n_i}\right), \tag{1}$$

where the weight of a term i in the document vector for j is the product of the frequency of term i in j and the log of its inverse document frequency in the data set, with n_i representing the number of documents in the data set that contains term i, and N representing the total number of documents in the data set.

Rather than only touching the surface of the document by counting the query terms, the proximity document-retrieval methods used spatial location information as a new factor to calculate the document score in information retrieval. The shortest substring retrieval model [12] was applied to the task of ranking documents. The document scores were based on the shortest substring of text in the document that matched the query. This was done by creating a data structure called a 'Generalized Concordance List' (GCL). These GCLs contain the span of a given term throughout the document. The method is limited when more than two query terms are used.

Based on the hypothesis that the closer the query terms are in the document, the more relevant the document is, [13] authors estimated the relevance of a document to a query by computing the fuzzy proximity degree of the query term occurrences in each document. This model is able to deal with Boolean queries but contrary to the traditional extensions of the basic Boolean model, it does not use a proximity operator explicitly. The drawback of this model is that simple queries based on the OR operator produce ranking scores that contradict the model's hypothesis.

The SBIRM is another model used in information retrieval systems to rank the documents according to document scores. Introduced by [8], this model combines the frequency and proximity of the query terms. The SBIRM is able to overcome the VSM limitation by comparing the query terms in their spectral domain rather than their spatial domain. The steps for ranking the documents according to the SBIRM are as follows:

1. The document is represented by a term signal for each query term.
2. The term signals are converted into term spectra using a spectral transform.
3. The spectral domain magnitude component shows the query term frequency.
4. The spectral domain phase component shows the query term position.
5. The components are summed to obtain the document score.

Based on SBIRM, relevant documents have a high magnitude and a common phase value across query term signals.

The DWT outperformed the Fourier transform and the DCT in document ranking and produced on average a 4% increase in precision when compared to the corresponding VSM [9]. Moreover, it is one of the frequently used transforms in data-mining applications. Wavelets [14] have many favorable properties, such as vanishing moments, multiresolution decomposition structure, and a wide variety of basic functions. These properties could provide considerably more efficient and effective solutions to many data-mining problems [15]. The DWT's advantage is its ability to break a signal into wavelets of different scales and positions, so that it can analyze the patterns of the terms in the document at various resolutions (whole, halves, quarters, or eighths). By contrast, the Fourier transform provides frequency information for the whole document, but does not enable researchers to focus on important portions of the document.

3. Arabic Chat Classification (ACC) Framework

Our proposed framework, Arabic Chat Classification (ACC), combines Information Retrieval and Machine Learning methods. With this hybrid approach, we can benefit from the document scores of the IR method to improve the performance of the classifier and minimize the wrong prediction class in Arabic text-based chat conversations.

The SBIRM has demonstrated its success in ranking documents in terms of precision [8]. Performing different transformations to convert term signals to the frequency domains allows us to use the spatial information in a more convenient and meaningful way.

ACC takes advantage of this spatial information by replacing the VSM with term signals that use DWT to give a more accurate performance of the classifier and increase the correct prediction class. The document scores that are obtained from the SBIRM are used as features of the NB algorithm.

3.1 ACC Framework Architecture

The ACC framework architecture is composed of two main phases, the Learning phase and the Predication phase. The two phases altogether consist of five steps: Text Preprocessing, Term Signal Formation, Term Transform (DWT), Feature Extraction, and the Naïve Bayes Algorithm. The ACC framework stages are shown in Fig.1. Within each step, a different method can be considered.

We discuss each of the ACC steps in detail, as it is essential to understand how ACC is implemented.

3.1.1 Text Preprocessing

Text preprocessing is an essential part of any text mining application, as the tokens at this stage are the fundamental units passed to all further processing stages. The chat-conversation text document in ACC must pass through the classical steps of tokenization, stopword removal, and stemming.

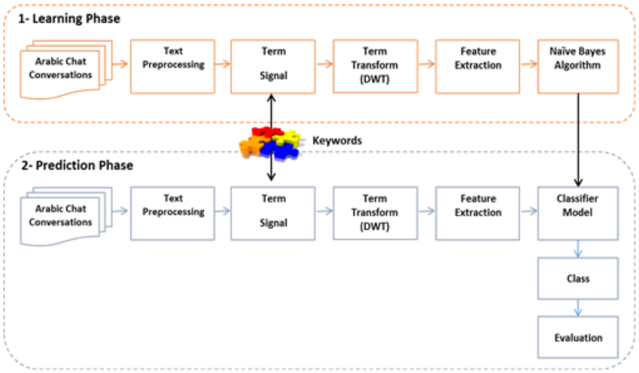

Fig. 1 ACC framework architecture

Tokenization

The first step of text preprocessing is the tokenization. Tokenization is the task of converting raw text files into a well-defined sequence of linguistically-meaningful units (tokens). The chat conversation-text document is split into a stream of words by removing all punctuation marks, brackets, hyphens, numbers, symbols, and non-Arabic words.

Stopword Removal

Removing stop words is another common step in text preprocessing. The stop words are the most frequently used and insignificant words, which are useless in information retrieval, and text mining. For Arabic, stop words include pronouns, prepositions, adverbs, days of the week, and months of the year. Stop words are removed because they do not help in determining a document's topic and also for dimensional redaction. The Arabic stopword list for this research has been adopted from [16].

Stemming

Word stemming in Arabic is the process of removing all of a word's prefixes and suffixes to produce the stem or the root [17]. The classification task applies a stemming process in text preprocessing because it makes the tasks less dependent on particular forms of words. It also reduces the size of the vocabulary, which might otherwise have to contain all possible ward forms.

Arabic stemming algorithms can be classified according to the level of analysis technique, as either stem-based or root-based algorithms. Stem-based algorithms remove prefixes and suffixes from Arabic words, while root-based algorithms extract the roots of the words. Light stemming refers to the process of stripping off a small set of prefixes and/or suffixes without attempting to find roots [18]. On the other hand, root-based algorithms reduce such Arabic words (المكتبة الكاتب الكتاب) which mean (the library), (the writer), and (the book) respectively, to one root (كتب), which means (write). In contrast, the light stem-based maps the word (الكتاب) which mean (the book) to (كتاب), (الكاتب) which mean (the writer) to (كاتب), and (المكتبة) which mean (the library) to (مكتبة).

The Alkhalil Arabic Morphology System, an open source tool, was used to find the roots and stems of the words. The system was developed in collaboration with the Arab League Educational Cultural and Scientific Organization and King Abdul Aziz City for Science and Technology[19].

3.1.2 Term Signal Formation

The term signal, introduced by [8], is a vector representation of terms that not only describes the frequencies of a term, but also its occurrence in particular partitions or bins within the document. The term signal shows how the term is spread throughout the document. To create a term signal, a document is first divided into a user-defined number of segments/bins B. Then, the term signal for term t in document d is represented by Eq. (2).

$$\tilde{f}_{d,t} = [f_{d,t,0}\, f_{d,t,1} \cdots\ f_{d,t,B-1}], \qquad (2)$$

where $f_{d,t,0}$ is the frequency of term t in the first bin of document d. $f_{d,t,b}$ can also be considered the b^{th} signal component of term signal $\tilde{f}_{d,t}$. For example, suppose that document d is divided into eight bins, B=8 (which is also the spectrum signal length), and consists of two terms, t_1 and t_2. Fig. 2 provides a visual example of how the term signals are obtained for terms t_1 and t_2.

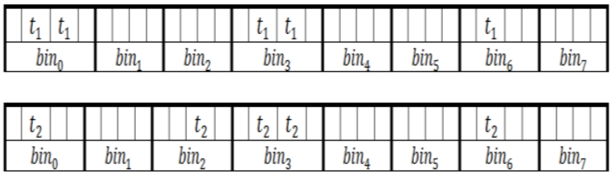

Fig. 2 A visual example of how the term signals are obtained.

As shown in Fig. 2, $t_1$occurs two times in bin_0, two times in bin_3, and one time in bin_6; t_2 occurs one time in bin_0, one time in bin_2, two times in bin_3, and one time in bin_6. The term signals for terms t_1 and t_2 are shown in Eq. (3).

$$\tilde{f}_{d,t_1} = [2\ 0\ 0\ 2\ 0\ 0\ 1\ 0] \qquad \tilde{f}_{d,t_2} = [1\ 0\ 1\ 2\ 0\ 0\ 1\ 0] \qquad (3)$$

The keyword terms are provided in this stage. These include a set of words known as topic keywords, which characterize a particular topic. For example, if a text document includes the keywords “information” and "computer," this document should belong to the IT category. If another text document included the keywords "information and technology," this document should belong to the IT category as well.

For each keyword term in keyword set, we match chat conversation tokens to create the term signals before performing preweighting.

Preweighting

This process involves applying weights before performing the Discrete Wavelet Transform. This might be completed to increase the accuracy of the positive class scores. We will use TF X IDF for preweighting. Since the keyword terms are divide into spatial bins, the modified version of the TF X IDF weighting scheme will be applied [10].

We apply weighting to each spatial bin to find term bin frequency × inverse document frequency (TF X IDF) Eq. (4).

$$\mathrm{TBF} = 1 + \log_e f_{d,t,b}\text{, and}$$
$$\mathrm{IDF} = \log_e\left(1 + \frac{N}{\mathrm{DocFreq}(t)}\right). \quad (4)$$

In the Eq. (4) formulas, $f_{d,t,b}$ is the count of term t in spatial bin b of document d, and IDF depends on the number of documents in which term t appears ($\mathrm{DocFreq}(t)$) and the number of documents (N) in the dataset. Thus, the weight signal $\tilde{\omega}_{d,t}$ for term t in the keyword set in document d is represented by Eq. (5).

$$\tilde{\omega}_{d,t} = [\omega_{d,t,0}\ \omega_{d,t,1} \cdots\ \omega_{d,t,8}]. \quad (5)$$

3.1.3 Term Transform (DWT)

We chose the Haar wavelet transform to provide us with the different levels of resolution of a signal. The Haar wavelet is equivalent to 1 cycle of a square wave, which has wavelet coefficients (high-pass filter) of $[\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}]$ and a scaling function (low-pass filter) $[\frac{1}{\sqrt{2}}\ \frac{1}{\sqrt{2}}]$. To perform the Discrete Wavelet Transform, we take every possible scaled and shifted version of the wavelet and, by finding the dot product, determine how much of this wavelet is within our signal [9]. The wavelet spectrum of the term signal $\tilde{\zeta}_{d,t}$ is defined as Eq. (6).

$$\tilde{\zeta}_{d,t} = [\zeta_{d,t,0}\ \zeta_{d,t,1} \cdots \zeta_{d,t,B-1}], \quad (6)$$

where $\zeta_{d,t,b} = H_{d,t,b}\exp(i\theta_{d,t,b})$ is the bth spectral component of the tth keyword term in the keyword set in the dth document with magnitude $H_{d,t,b}$ and phase $\theta_{d,t,b}$.

3.1.4 Feature Extraction

The document scores [9], which are the features of the classifier, are obtained as follows. The magnitude $H_{d,t,b}$ vector is defined in Eq. (7).

$$H_{d,t,b} = |\zeta_{d,t,b}|, \quad (7)$$

The phase vector $\phi_{d,t,b}$ is defined in Eq. (8).

$$\phi_{d,t,b} = \frac{\zeta_{d,t,b}}{|\zeta_{d,t,b}|} = e^{(i\theta_{d,t,b})}. \quad (8)$$

The Haar wavelet transform does not produce complex values when applied to a real signal. Hence, the sign of the component is treated as the phase. Therefore, $\zeta_{d,t,b}$ is real and $\theta_{d,t,b}$ must be of the form πn, where n is an integer. This implies that we will have only $\phi_{d,t,b} \in \{-1,1\}$. The zero-phase precision formula can be simplified to Eq. (9).

$$\text{Zero Phase Precision} \coloneqq \bar{\Phi}_{d,b} = \left|\frac{\sum_{t\in K}\mathrm{sgn}(\zeta_{d,t,b})}{\#K}\right|, \quad (9)$$

where K is the set of keyword terms, #K is the cardinality of the set K, and sgn is defined in Eq. (10).

$$\mathrm{sgn}(\zeta_{d,t,b}) = \begin{cases} 1 & \text{if } y \geq 0 \\ 0 & \text{if } y = 0, \\ -1 & \text{if } y < 0 \end{cases} \quad (10)$$

To obtain the spectral component score, apply Eq. (11).

$$s_{d,b} = \bar{\Phi}_{d,b}\sum_{t\epsilon k} H_{d,t,b}, \quad (11)$$

and to combine the document score, use Eq. (12).

$$S_d = \|\tilde{s}_d\|_p, \quad (12)$$

where $\tilde{s}_d = [s_{d,0}\ s_{d,1} \cdots s_{d,B-1}]$ and $\|\tilde{s}_d\|_p$ is the l^p norm given by Eq. (13).

$$\|\tilde{s}_d\|_p = \sum_{b=0}^{B-1}|s_{d,b}|^p \quad (13)$$

In our experiments, we will be examining S_d for p =2.

3.1.5 Naïve Bayes Algorithm

The Naïve Bayes algorithm was chosen as a classifier. It is a simple, straightforward, frequently used method for supervised learning. It is considered one of the top 10 data mining algorithms [20], and it is based on the Bayesian theorem with strong independence assumptions. Despite its simplicity, NB can often outperform more sophisticated classification methods. It can predict class membership probabilities such as the probability that a given sample belongs to a particular class. Using the Bayes theorem, the probability of a document d being in class C_i is calculated in Eq. (14) as:

$$\mathbf{P}(\mathbf{C_i}|\mathbf{d}) = \frac{P(C_i)P(d|C_i)}{P(d)}, \quad (14)$$

where $P(C_i|d)$, $P(C_i)$, $P(d|C_i)$, and $P(d)$ are called the posterior probability, prior probability, likelihood, and evidence respectively.

Let T be a training set of samples, each with their class labels. There are i classes: C_1 , C_2 ,…, C_i. Each sample is represented by an n-dimensional vector, X = $\{x_1$, x_2,…, $x_n\}$, showing n measured values of the n attributes, A_1, A_2 ,…, A_n , respectively. To classify a new document X, NB calculates posteriors for each class, and assigns the document to that class C_i, for which it achieves the highest posterior probability. As $P(d)$ is the same for all classes, only $P(C_i)P(d|C_i)$ need to be maximized. Thus, the best class in the NB classification is the most likely or maximum a posteriori (MAP) class, as described in Eq. (15).

$$C_{map}: P(C_i|X) \approx P(C_i)\prod_{k=1}^{n} P(x_k|C_i) \quad (15)$$

The Naïve Bayes classifier has different models / variants [21]. Since our features are continuous values (document scores), then we typically assume that the values have a normal distribution with a mean μ and standard deviation σ, as defined by Eq. (16).

$$g(X, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \quad (16)$$

Thus, $P(x_k|C_i) = g(x_k, \mu C_i, \sigma C_i)$, where μC_i and σ C_i are the mean and standard deviation of attribute values x_k of the training set documents for class C_i, respectively. The prior probabilities of the class $P(C_i)$ may be estimated as the relative frequency of class C_i , in training set $P(C_i) = \frac{N_{C_i}}{N}$, where N_{C_i} is the number of documents in class C_i while N is the total number of documents. We may conclude that while we have one attribute, which is the document score S_d the posterior probability of the document d can be calculated by using Eq. (17).

$$P(C_i|d) \approx P(C_i) * g(d, \mu C_i, \sigma C_i)$$

$$P(C_i|d) \approx P(C_i) * \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(S_d-\mu)^2}{2\sigma^2}}, \quad (17)$$

where μC_i and σ C_i are the mean and standard deviation of S_d values of training set documents for class C_i , respectively.

3.2 The ACC Framework Algorithms

3.2.1 Learning Phase Algorithm

In the learning phase, the ACC framework uses Algorithm 1.

ALGORITHM 1. Learning Phase

```
1. For each document d in training set
        2. For each keyword term in keyword set t ∈ K
            - Create term signals $\tilde{f}_{d,t}$
            - Weight signals $\tilde{\omega}_{d,t} = \omega(\tilde{f}_{d,t})$
            - Transform signals $\tilde{\zeta}_{d,t} = DWT(\tilde{\omega}_{d,t})$
            3. For each spectral component $\zeta_{d,t,b} \in \tilde{\zeta}_{d,t}$
                - Calculate signal component magnitudes $H_{d,t,b} = |\zeta_{d,t,b}|$
                - Calculate signal component phase $\phi_{d,t,b} = sgn(\zeta_{d,t,b})$
                - For each component b in term signal spectrums
                    a. Calculate the zero phase precision $\bar{\Phi}_{d,b} = \left|\frac{\sum_{t\epsilon K} sgn(\zeta_{d,t,b})}{\#K}\right|$
                    b. Obtain component score $s_{d,b} = \bar{\Phi}_{d,b}\sum_{t\epsilon k} H_{d,t,b}$
            4. Combine component scores to obtain document score $S_d = \sum_{b=0}^{B-1}|s_{d,b}|^p$
5. For each class $C_i$
    - Calculate the mean $\mu C_i$ of attribute values $S_d$ of training set documents
```

```
        - Calculate the standard deviation $\sigma\ C_i$ of
          attribute values $S_d$ of training set documents
        - Calculate the prior probabilities of the class $C_i$
          $P(C_i)$
```

3.2.2 Prediction Phase Algorithm

In the prediction phase, the ACC framework uses Algorithm 2.

.

ALGORITHM 2. Prediction Phase

```
1. For each document d in test set
    2. For each keyword term in keyword set $t \in K$
        - Create term signals $\tilde{f}_{d,t}$
        - Weight signals $\tilde{\omega}_{d,t} = \omega(\tilde{f}_{d,t})$
        - Transform signals $\tilde{\zeta}_{d,t} = DWT(\tilde{\omega}_{d,t})$
        3. For each spectral component $\zeta_{d,t,b} \in \tilde{\zeta}_{d,t}$
            - Calculate signal component magnitudes $H_{d,t,b} = |\zeta_{d,t,b}|$
            - Calculate signal component phase $\phi_{d,t,b} = sgn(\zeta_{d,t,b})$
            - For each component b in term signal spectrums
                c. Calculate the zero phase precision $\bar{\Phi}_{d,b} = \left|\frac{\sum_{t \in K} sgn(\zeta_{d,t,b})}{\#K}\right|$
                d. Obtain component score $s_{d,b} = \bar{\Phi}_{d,b} \sum_{t \in k} H_{d,t,b}$
        4. Combine component scores to obtain document score $S_d = \sum_{b=0}^{B-1} |s_{d,b}|^p$
5. For each class $C_i$
    - Calculate $g(d, \mu C_i, \sigma C_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(S_d-\mu)^2}{2\sigma^2}}$
    - Calculate posterior probability $P(C_i|d) \approx P(C_i) * g(d, \mu C_i, \sigma C_i)$
    - Choose class that maximizes $P(C_i|d)$
```

4. Experiments and Results

The ACC was programmed in Java. To examine the performance of the framework, two experiments were conducted using the data set. The first experiment was for root-based chat conversations, and the second experiment was for stem-based chat conversations.

4.1 Data Set and Evaluation Method

To conduct the experiments, the topic that should be detected in chat conversations must be identified. “Arab Spring Revolution” was chosen as the illegal chat conversation topic. The data set was collected manually from the Internet, and consisted of 100 Arabic, text-based, chat conversation files divided into:

- 34 files belonging to the illegal class, the “Arab Spring Revolution” topic;
- 66 files belonging to the legal class, which consisted of another topic.

To evaluate the ACC framework, the 10-fold cross validation method was used [22], [23]. The entire data set was divided into 10 equal folds. Fold numbers 1, 2, 3, and 4 had 4 text files that belonged to the illegal class and 6 text files that belonged to the legal class, and fold numbers 5, 6, 7, 8, 9, and 10 had 3 text files that belonged to the illegal class and 7 text files that belonged to the legal class. The evaluation metrics were accuracy, precision, recall and F-measure.

4.2 Keyword Set

The keyword set includes keywords that characterize a particular topic. It was determined from observation of the collected data set. The list of the keywords that belonged to the “Arab Spring Revolution” topic and their meanings are shown in Table 1.

Table 1: Keyword terms with their English meaning

Arabic Keyword	English Meaning	Arabic Keyword	English Meaning
ثورة	Revolution	حكومة	Government
مظاهرة	Demonstration	حرب	War
سلاح	Weapon	شهداء	Martyrs
اشتباكات	Clashes	نظام	System
انشقاقات	Splits	القاعدة	Kaida
احتجاج	Protest	فاسد	Corruption
جيش	Army	طائفية	Sectarianism
تفجير	Bombing	حرية	Freedom
رصاص	Bullet	مقاومة	Resistance
عسكري	Military	اليمن	Yemen
دستور	Constitution	سوريا	Syria
شعب	People	البحرين	Bahrain
فتنة	Mesmerized	ليبيا	Libya
اعتداء	Abuse	مصر	Egypt

4.3 ACC Framework Experiments' Results

Table 2 shows the ACC framework's performance results for the root-based chat conversations and the stem-based chat conversations. As shown in Table 2, the ACC framework experiment for the root-based chat conversations outperformed the stem-based chat conversations in accuracy, precision, and F-measure. Both experiments produced the same recall rate of 91.18%. Thus, the text preprocessing (i.e., stemming techniques) influenced the ACC framework's performance.

Table 2: ACC framework's performance results

	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Root-Based	**94**	**91.18**	**91.18**	**91.18**
Stem-Based	**91**	**83.78**	**91.18**	**87.32**

To compare our ACC framework's performance results with the VSM, another two experiments were conducted using the RapidMiner Studio tool. The experiments were run with the same data set and the same evaluation method, and the classification algorithm was NB using the TBF X IDF formula.

Table 3 shows the VSM's performance results for both the root-based and stem-based chat conversations.

Table 3: VSM performance results

	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Root-Based	**64**	**25**	**2.94**	**5.26**
Stem-Based	**63**	**33**	**8.82**	**13.92**

The VSM performance results were worse than those of our ACC framework. The main reasons for this were the small data set and the fact that no keywords had been specified. We may conclude that our hybrid approach (i.e., the ACC framework) produces high performance in small data sets, using a root-based stemming technique for classifying Arabic text-based chat conversations.

5. Conclusions

Chat conversations create a new challenge for knowledge discovery due to their structure. In this research, the ACC framework that combines the features of information retrieval and the power of machine learning systems was developed to classify topics of interest in Arabic text-based chat conversations. The ACC framework enabled the acquisition of high-performance classification results, and valuable information from chat conversations was revealed.

Using the Spectral-Based Information Retrieval Method, Arabic text-based chat conversations were classified through the transfer of the text from the spatial domain to the spectral domain, which increased the classifier performance measures.

The Arabic Chat Classification (ACC) framework was developed and implemented using the Java programming language. It was composed of two main phases: the Learning phase and the Predication phase. The two phases consist of five steps: Text Preprocessing, Term Signal Formation, Term Transform (DWT), Feature Extraction, and the Naïve Bayes Algorithm. Within each step, a different method was applied. The ACC framework was tested and evaluated, with experiment results showing that the root-based chat conversations outperformed the stem-based chat conversations in accuracy, precision, and F-measure. Both experiments produced the same recall result rate of 91.18%. Thus, the text preprocessing (i.e. stemming techniques) influenced the ACC framework's performance. Using root-based stemming technique in classifying Arabic text-based chat conversations provided high performance results.

Many areas for further exploration exist. A good starting point for future research may include broadening the framework by adding different signal transformations, and conducting comparisons among them. A framework using different classification algorithms can also be developed and tested using Arabic text-based chat conversations. This framework may also be useful for different applications, such as document or web page classifications. In addition, building a tool to extract from Arabic text-based chat conversations the keywords that describe a particular topic may be valuable. Finally, semiotics can be used to detect illegal activities in Arabic text-based chat conversations.

References

[1] Adams P., and Martell C., "Topic Detection and Extraction in Chat," *IEEE International Conference on Semantic Computing*; Santa Clara, CA, pp. 581–588, 2008

[2] Bengel J. et al., "ChatTrack : Chat Room Topic Detection Using Classification," *2nd Symposium on Intelligence and Security Informatics*, Tucson, Arizona, pp. 266–277, 2004.

[3] Dong H., Siu H., and Yulan H., "Structural Analysis of Chat Messages for Topic Detection," *Online Information Review*, vol.30, no. 5, pp. 496–516, 2006.

[4] Elnahrawy E., "Log-Based Chat Room Monitoring Using Text Categorization : A Comparative Study," The IASTED International Conference on Information and Knowledge Sharing, US Virgin Islands, pp.111-115 , 2002.

[5] Hui S., He Y., and Dong H., "Text Mining for Chat Message Analysis," *IEEE Conference on Cybernetics and Intelligent Systems*, pp. 411–416, 2008.

[6] Özyurt Ö., and Köse C., "Chat Mining: Automatically Determination of Chat Conversations' Topic in Turkish Text

Based Chat Mediums," *Expert Systems with Applications*, vol. 37, no. 12,pp. 8705–8710, 2010.

[7] Salton G., Wong A., and Yang C., "A Vector Space Model for Automatic Indexing." *ACM* vol.18, no. 11,pp. 613–620, 1975.

[8] Park L ., "Spectral Based Information Retrieval," The University of Melbourne, 2003.

[9] Park L. , Ramamohanarao K., and Palaniswami M., "A novel document retrieval method using the discrete wavelet transform," *ACM Transactions on Information Systems*, vol. 23, no. 3, pp. 267–298, 2005.

[10] Park L. , Ramamohanarao K., and Palaniswami M., "Fourier Domain Scoring : A Novel Document Ranking Method," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 16, no. 5, pp. 529–539, 2004.

[11] Park L. , Ramamohanarao K., and Palaniswami M., "A novel document ranking method using the discrete cosine transform," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 1, pp. 130–5, 2005.

[12] Clarke C. and Cormack G., "Shortest-substring retrieval and ranking," *ACM Transactions on Information Systems*, vol. 18, no. 1, pp. 44–78, 2000.

[13] Beigbeder M. and Mercier A., "An Information Retrieval Model Using the Fuzzy Proximity Degree of Term Occurences," *Proceedings of the 2005 ACM symposium on Applied computing*, pp. 1018–1022, 2005.

[14] Walker J., *A Primer on Wavelets and Their Scientific Applications*, Chapman and Hall/CRC, Wisconsin, 2008.

[15] Li T. et al., "A survey on wavelet applications in data mining," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 49–68, 2002.

[16] "Arabic Stop words," 2010. [Online]. Available: http://sourceforge.net/projects/arabicstopwords. [Accessed: 11-May-2013].

[17] Khoja S. , "APT : Arabic Part-of-speech Tagger," *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 20–26, 2001.

[18] Al-Sughaiyer I. and Al-Kharashi I. , "Arabic morphological analysis techniques: A comprehensive survey," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 189–213, 2004.

[19] Boudlal A. et al., "Alkhalil Morpho Sys 1 : A Morphosyntactic analysis system for Arabic texts," *Proceedings of ACIT'2010*, pp. 1–6, 2011.

[20] Wu X. et al., *Top 10 algorithms in data mining*, vol. 14, no. 1. Springer-Verlag, pp. 1–37, 2007.

[21] Al-aidaroos k., Bakar A., and Othman Z., "Naïve Bayes Variants in Classification Learning," Proceeding of the International conference on Information Retrival and Knoledge Management, Selangor, pp. 276–281, 2010.

[22] Forman G. and Scholz M., "Apples-to-Apples in Cross-Validation Studies : Pitfalls in Classifier Performance Measurement," *ACM SIGKDD Explorations*, vol. 12, no. 1, pp. 49–57, 2010.

[23] Hsu C., Chang C., and Lin C., "A Practical Guide to Support Vector Classification," vol. 1, no. 1, pp. 1–16, 2010.

Arwa Diwal has received a MSc degree from King Abdul Aziz University in Jeddah. She is working as Lecturer in the Information Systems Department, Faculty of Computing and Information Technology.

Dr. Mahmod Kamel received a PhD degree from Al-Azhar University, Egypt. At Present, he is an Assistant Professor in the Information Systems Department, King Abdul Aziz University. He has more than 12 years of experience in the field of teaching and training in the educational sector. He has been guiding Master students for the past 8 years.

Dr. Mohmmed Dahab received a PhD in Computer Science from Cairo University, Egypt, in 2007. Currently, he is an Assistant Professor at King Abdul Aziz University. His research interests include applications of Expert Systems, Natural Language Processing, Text Mining, Information Retrieval, and Machine Learning.