APPLYING DISTRIBUTIONAL SEMANTICS TO ENHANCE CLASSIFYING EMOTIONS IN ARABIC TWEETS

Shahd Alharbi[1] and Matthew Purver[2]

[1]Department of Software Engineering, King Saud University, Riyadh, Saudi Arabia
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

ABSTRACT

Most of the recent researches have been carried out to analyse sentiment and emotions found in English texts, where few studies have been conducted on Arabic contents, which have been focused on analysing the sentiment as positive and negative, instead of the different emotions' classes. Therefore this paper has focused on analysing different six emotions' classes in Arabic contents, especially Arabic tweets which have unstructured nature that make it challenging task compared to the formal structured contents found in Arabic journals and books. On the other hand, the recent developments in the distributional sematic models, have encouraged testing the effect of the distributional measures on the classification process, which was not investigated by any other classification-related studies for analysing Arabic texts. As a result, the model has successfully improved the average accuracy to more than 86% using Support Vector Machine (SVM) compared to the different sentiments and emotions studies for classifying Arabic texts through the developed semi-supervised approach which has employed the contextual and the co-occurrence information from a large amount of unlabelled dataset. In addition to the different remarkable achieved results, the model has recorded a high average accuracy, 85.30%, after removing the labels from the unlabelled contextual information which was used in the labelled dataset during the classification process. Moreover, due to the unstructured nature of Twitter contents, a general set of pre-processing techniques for Arabic texts was found which has resulted in increasing the accuracy of the six emotions' classes to 85.95% while employing the contextual information from the unlabelled dataset.

KEYWORDS

SVM, DSM, classifying, Arabic tweets, hashtags, emoticons, NLP &co-occurrence matrix.

1. INTRODUCTION

The recent dominance of the electronic social media websites and especially Twitter which has become the major form of communication and expressing peoples' emotions and attitudes in the world generally and Arab world especially, has considered the significant encouraging factor for carrying this study. According to[1] which has published 2014 statistics for Twitter usage in Arab

Dhinaharan Nagamalai et al. (Eds) : NATL, CSEA, DMDBS, Fuzzy, ITCON, NSEC, COMIT - 2018
pp. 15–34, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80602

world which showed increasing the number of active Twitter users in Arab world to more than 5 million. Additional remarkable statistic shows that the average number of tweets in 2014 written using Arabic language account over than 75% among other languages, which considered as a strong proof for the Arabic language as being one of the fastest growing languages on the web.

An additional encouraging factor, was based on the recent features and services provided by Twitter to its public users, such as streaming APIs supported by language mode which facilitate the automatic collection of a huge amount of its data without manual interventions, which has raised Twitter value as being a wealth source, gold mine and major area of interest to different researchers in the subjectivity analysis field to mainly analyse texts based on being subjective or objective.

Special area has been explored by researchers to analyse users' positive and negative sentiments and mine their opinions and reviews from unusual, and informal short texts which differ from the formal texts found in newspapers and documents.

Many of these researches have been applied to business tasks for companies to improve their services and products based on analysing customers' feedback. Similarly, customers can use analysed reviews to reveal insights about services and products. In addition, many other applications have been seen in determining public's attitudes and views with respect to special topic or incident.

The remaining sections of this paper is organised into a number of different sections, section 2 highlights some of the previous studies with their developed models in the different areas related to the distributional semantics and to the sentiment and emotions analysis; section 3 represents the methodology followed in building the proposed model for enhancing the automatic classification of the different emotions found in Arabic tweets, which was developed by [2]; in addition, section 4 describes some of the different carried experiments with their analysed results, and the final sections represents the conclusions drawn from the previously performed experiments as well as some of the faced limitations and the expected future work to eliminate these limitations.

2. RELATED WORK

2.1. Sentiment and Emotion Analysis in Arabic Texts

Lately the growth of the number of Arab social network users, produce a massive amount of Arabic texts and reviews available through the social medium which highlighted the need for automatic sentiment and emotions identification from this large generated reviews and texts.

Although some emotions and sentiment analysis studies have been conducted to non-English languages, most of these studies have been focused on non-Arabic languages. However, a sentiment analysis study was done by [3] to classify opinions in a number of web review forums, which was tested on Arabic dataset, through developing entropy weighted genetic algorithm, EWGA, for feature selection in addition to the different features extraction components that were used to compute Arabic features' linguistic characteristics. Results have proved the effectiveness of the built methodology for analysing and classifying sentiment in different languages through the use of SVM.

A former study was done by [4] which developed a tool dedicated for analysing colloquial Arabic texts that appear in the web forums and social media websites. A limitation of the dependency on human interventions and judgments to overcome the problems generated form the non-standardised colloquial Arabic texts has been resolved by their proposed tool through a game-based lexicon which is based on human expertise to classify the sentiment of the phrases.

Additional methodology was used for classifying sentiment, based on the subjectivity of the different words in the constructed lexicon of the words resulted from phrases segmentation.

It has been observed that many sentiment analysis studies have been conducted on analysing large texts and documents, however, few years after carrying out [4] study, [5] have analysed sentiments based on the sentimental majority in Arabic sentences, through calculating the number of positive and negative phrases and classify the sentence according to the dominant sentiment which recorded an accuracy of 60.5%. Their approach was based on their previously followed game-based approach to annotate large cuprous manually.

Moreover, a model has been introduced by [6], known as SAMAR, which was designed to recognize the subjectivity and to identify sentiment for sentence-level Arabic texts in the social media websites' content such as tweets. Compared to [4], SAMAR has the ability to analyse sentiment found in MSA and colloquial Arabic phrases. To overcome the difficulty and the inaccuracy resulted from classifying sentiment in colloquial Arabic phrases that vary in the rules and vocabularies compared to MSA, analysed Twitter messages were annotated manually according to both, their subjectivity, as well as the different Arabic texts' classes as MSA and colloquial.

Despite the long success journey of analysing the sentiment and the subjectivity in Arabic texts, a lack of identifying emotions in Arabic contents of the social media was found which can lead to an inaccurate classification results for analysing emotion-based texts.

There was a remarkable solution developed by [2] which overcome this analysis limitation through classifying emotions found in Arabic tweets. This recent developed tool has the ability to detect the dialects found in Arabic tweets as well as to identify the different emotions used by [7]and others, through using SVM classification algorithm with n-gram order features and distant supervision approach similar to [7], [8] and [9] approaches. Results have recorded a good performance in predicting emotions for the different conventional markers in Arabic tweets according to the different emotions based on the human judgments, these results were between 93% and 74% and between 81% and 72% for emoticons and hashtags respectively.

2.2 Distributional Semantics

It has been observed from the different studies in developing Distributional Semantic Model (DSM) approaches, the strong dependency on the assumption that the similarity of meanings between linguistic entities in any textual data can be defined in terms of the distributional properties of these linguistic entities. Therefore, the notion of utilizing these distributional properties for inferring meanings considered the hallmark of any developed DSM, which has come to be known as distributional hypothesis that considered the main idea behind the distributional semantics, therefore DSM need to be implemented on its basis. Rubenstien and Good enough have reported on 1965 the early theory of distributional hypothesis in which words that have similar meanings occur in similar contexts. Based on their hypothesis, the predicted

similarities of the different words were evaluated based on their correlation with the predefined 65 noun pairs synonyms' similarities supported through the human judgments.

In the past four decades, a distributional model was developed by Harris who considered as the basic motivation for the distributional hypothesis where the differences in semantics between words can be correlated with the differences in distributions between these words.

Similar distributional hypothesis was defined by Schutzeand Pedersen on1995, such that different words will occur with similar neighbours if they have similar meanings based on the existence of sufficient text materials. Where the idea of the distributional semantics hypothesis defined by [10]is based on locating words with similar distributional properties in similar regions of the word-space.

Another hypothesis was defined by [11] model for testing the efficiency of the high order co-occurrence for detecting similarities. [12]have developed their approach based on their assumption that similarities in syntactic structure can result in semantic similarities.

Moreover, [13] model have depended on analysing Arabic texts similarities to identify the different participants groups presented in the online discussion based on the hypothesis that participants can share the same opinion in a discussion if they focus in the similar aspects of the discussion topic.

Despite the different distributional hypothesises which have been defined in different studies, they all fall under the idea that two words are expected to be semantically similar if they have similar co-occurrences in the observed context.

In addition to the distributional hypothesis, word-space vectors represents another fundamental basis for developing DSM approaches. In the past two decades, [14]found that DSM is based on the similarities and the distances between words in the vector space, in which semantically similar words are close, since semantics are expressed by the same set of words.

Moreover, [15] have found that word-space models utilises distributional patterns of phrases and words collected from large textual data to represent semantic similarities through the proximity between different phrases or words in an n-dimensional word-space.

Basically, two different approaches were recognized for the different developed models of DSM. The first methodology followed by researchers is based on the word-space vectors for building distributional profiles based on the surrounding words of the different terms as demonstrated by [14]. The second approach, known as latent semantic analysis model, LSA, which is one of the first applications of DSM and word-space vectors which are used in information and documents retrieval. This model is based on building distributional profiles of the words' occurrence based on the word-by-document matrix. [16]have introduced LSA approach using a word-by-document co-occurrence matrix in which contexts represented are based on the different documents in which words appear.

Despite the different types of distributional resources used for inferring similarities in these approaches, they share their main objective of inducing knowledge about meanings and similarities indirectly from co-occurrence of large textual data [10].

2.3. Semantic Similarity Detection in Arabic

It has been emerged a considerable number of studies for classifying Arabic texts and documents based on their similarities using distributional semantics and statistical properties with different similarity measures. This area has been seen as a successful application in the IR, text summarizing, document clustering, questions answering and other domains.

Based on [13] and their previously identified hypothesis, an approach has been developed to detect subjectivity of the Arabic discussions and to identify discussions' topics based on similarities between participants' opinions in online Arabic discussion groups. Similarities were detected using word-vector spaces of 100 dimension for representing different participants' discussions which contain distributional measures for the 100 different participated topics. Results have showed that the word-vector representations considered a rich representation for explicitly illustrating different discussions' topics. It has been observed the contribution of the topic representations and the distributional statistics on enhancing the accuracy of identifying subjectivity of the different Arabic discussions along with their topics.

According to [17], cosine-based similarity has considered one of the well-known similarity measures applied to the documents in different applications as IR, NLP, machine translation and text mining for Arabic documents classifications. [17]have found the effectiveness of using cosine similarity measures in Arabic document classification tasks compared to other distance-based and similarity measures.

Prior to [17], [18] have conducted a study which aims to find the optimal classification model for classifying Arabic texts. To find the best coefficient for the vector-space model to classify Arabic documents, these documents were represented as vectors to compare between the different coefficients, as cosine. Results have found the superiority of the cosine measure compared to other coefficients. Moreover, the effectiveness of applying cosine measure in classifying Arabic texts was compared to other similarity measures as Naïve Bayes which outperforms the used cosine-based similarity measure. [18]have suggested combining different classification models which can increase the classification accuracy for Arabic text, as Naïve Bayes with cosine measure to determine similarity in the classified text. Following [18], [19] have investigated determining the similarity between Arabic texts using different bigrams techniques, word-based model, document-level model and vector-based model which can reflect the importance of the bigrams in the document, based on the weighted vectors of the represented document. Cosine similarity measure has been applied to the vector-based model for measuring similarities between tow vectors. On the other hand, word-based and document-level models have been used with different similarity measures as investigated by [20]for measuring similarities. Compared to the human similarity judgments, results have demonstrated the efficiency of the cosine similarity measures to determine similarities between Arabic texts and documents.

Although it can be observed form this set of highlighted studies the lack of classifying Arabic social media's texts based on applying different similarity measures, one of the most recent similarity and categorisation related studies was based on Arabic social media's texts especially Twitter posts, that was carried out by [21]. This study has delivered the problem of non-related results retrieved from searching Arabic tweets through developing a machine-learning model for summarising Arabic Twitter posts and especially Egyptian dialect posts which used cosine-similarity model as one of the developed approaches. A high performance was achieved by their proposed system compared to other summarisation algorithms.

2.4 Classification and Annotation

2.4.1 Arabic Text Classification

A considerable number of Arabic texts classification studies have used a similar classification methodology which was used in [7], [8] and [22] tools for classifying Arabic sentiments and emotions found in the different tweets.

Following [7], [2] have developed a tool which was built using SVM classifier for detecting and classifying emotions and dialects in Arabic tweets based on the same conventional markers and emotions' classes used by [7]. It has been proven that following[7] and [22] approaches of using distant supervision classification approach to automatically label emotions and dialects in Arabic tweets, can achieve more reliable results for emotions and dialects classification accuracies. Results have showed that changes in the classifier accuracies achieved, were based on the application domain (emotion or dialect detection). This has supported [23] findings regarding the reliance of the machine-leaning classifiers' performance on the domain applications used during the training process.

A study was carried out by [24] which encouraged the use of SVM in the area of classifying and categorizing Arabic texts, consequently, a comparative study was conducted by [25] for comparing between two machine-learning classification approaches including SVM for categorising Arabic texts using a large number of training and testing articles. Results have showed the role of features set size on the SVM performance, as it has been observed that larger set of features has increased SVM classification outcomes. A similar classifiers' comparison based study was done by [26] based on applying SVM and other traditional machine-learning classification techniques as Naïve Bayes classifier in classifying Arabic texts. Results have supported [25] findings, in which a better performance from SVM classifiers was generated when increasing the features set size, due to its ability to deal with sparse vectors of the classified documents. Moreover, the high dimensions of the vector-space represented from classified dataset enable SVM in handling large number of features.

Alternatively, many remarkable Arabic text classification studies have used different classification techniques. [27]has compared the efficiency of using distance-based and different machine-learning techniques as Naïve Bayes to classify large Arabic documents represented using word-vector spaces with their frequencies. Results have showed that Naïve Bayes machine-learning model has outperforms the distance-based model for classifying Arabic texts.

2.5 Arabic Text Pre-processing Approaches

Based on the different Arabic language specifications, many Arabic text classification studies have investigated the effect of the language properties on the trained and the classified texts. A study was carried by [28] which explored the effect of different Arabic pre-processing techniques on classifying texts by applying different term weighting and stemming approaches. According to [28] experiments, result have showed the superiority of SVM to classify Arabic processed texts compared to other text classifiers, moreover it has been found that the light stemming, considered the best feature reduction technique . Similarly, [29] have investigated the impact of using stemming in the Arabic text pre-processing. Results have showed that using stemming, SVM accuracy was increased compared to Naïve Bayes classifier. Moreover, [30] model has pre-

processed the review dataset using different approaches to remove spammed, noisy and duplicated reviews to avoid inconsistency during processing reviews and to guarantee a unique dataset contents which was used in the three different dictionaries, Arabic, English and emotions. Arabic opinion analysis model proposed by [31], has performed different pre-processing schemas as removing punctuations, symbols, digits from dataset, applying tokenisation as well as filtering non-Arabic texts to normalize analysed opinions. Recent development of Arabic text classification that was built by [2]for classifying Arabic emotions and dialects in Twitter messages has investigated the impact of applying similar pre-processing techniques that was followed by [28]. In addition to the different pre-processing schemas [2] and [28] have investigated the impact of normalising Arabic text as well as removing stop words as a feature reduction. Results examined by [2] have showed the correlation between SVM classification accuracy and different pre-processing techniques applied to classify emotions in Arabic tweets.

On the other hand, [32] have pre-processed Arabic documents that were used in the keyword extraction system to reduce features by portioning documents into sentences and removing the non-candidate keywords. [26]model has used SVM to classify Arabic texts with features' selection schemas during pre-processing steps. Stemming and eliminating stop words were used to reduce dimensionality. Results showed the positive effect of the light stemming on the Arabic text classification.

3. METHODOLOGY

The methodological approach followed in this study was based on applying the DSM measures from the unlabelled collected random dataset to build the co-occurrence matrix which was used to derive features that encode meaning and similarity. These features were then added into the standard n-gram features' order representation used in [2] classification approach in order to prove our early stated hypotheses for their ability in enhancing [2] model for classifying emotions found in Arabic tweets. The process followed was mainly divided into four stages, preparing labelled (keyword) and unlabelled (random) datasets, measuring co-occurrences for unlabelled dataset terms, preparing the best pre-processing techniques which was applied in [2] model and building the feature vectors for our classified datasets. For classification improvement purposes, the outcome of the latter stage was then used with the SVM classifier model developed and used by [2].

3.1 Preparing the labelled and unlabeled datasets

Two different Arabic Twitter datasets, labelled and unlabelled datasets, were collected automatically using some available Java libraries, Twitter4j. Streaming API was the main Twitter API used to access a global stream of random Arabic tweets.

The collected labelled dataset was used during the training and the testing stages of the classification process, with the different conventional markers, hashtags and emoticons, in the collected posts for the different six emotions' classes, happy, sad, anger, fear surprise and disgust. Where the collected unlabelled dataset was used for building the co-occurrence matrix which represented the contextual information. In both datasets a language request parameter was set to Arabic for restricting the collected data. Both datasets were further processed to eliminate duplicate and retweet data used during the classification process to avoid bias results in both datasets. Additional process included for the labelled dataset to eliminate tweets which include

mixed labels from different emotions' classes, to avoid confusing the classifier during the classification tasks with mixed labels in the training and testing sets. This features' reduction process applied at early stages considered a required step which was performed before classifying the different emotions. On the other hand, pre-processing techniques considered an optional features' reduction techniques which were applied when required.

3.2 Measuring co-occurrences for the unlabelled dataset

As DSM is based on obtaining distributional information in a high dimensional vectors, the proposed model represents these distributional information in RxC matrix to capture the co-occurrence frequencies measures. Although, it has been proven from previous studies the ability of the co-occurrence statistics to provide semantic information, the proposed approach aims to reveal an extension for its ability to strengthen the classification job by exploiting these measures.

3.3 Pre-processing Arabic Text

According to the Arabic text classification studies which were highlighted earlier, Arabic text-pre-processing techniques have proved their effective impact on the classification task. In our model this effect have a special importance during the feature reduction process due to the noisy and unstructured nature of the Arabic texts produced in Twitter by different users which makes similar messages undistinguishable by the classifier due to the differences in their presentation. Consequently, the proposed model has applied some optional techniques to normalise Arabic tweets for the classifier to simplify its job and therefor better results can be achieved. Although there was an absence for a standard defined set of techniques for pre- processing Arabic texts, our model has applied [2] approach using six different techniques during this optional feature reduction step such as removing the Arabic stop words, stemming, normalisation, removing diacritics as well as lengthening characters and reducing repeated characters. Unlike [2] during this processing stage our model has applied their techniques on the vocabularies-level for co-occurrence features as well as the tweets-level for the labelled dataset.

Choosing the best number of co-occurrence dimensions was based on the best average classification accuracy for the different emotions resulted from the different tested dimensions. Similarly, the best set of pre-processing techniques was selected based on the highest accuracy average resulted from classifying the six emotions labelled using hashtags and emoticons.

3.4 Building Features' Vectors

Apparently, co-occurrence statistics and distributional measures obtained from the co-occurrence matrix solely cannot considered as sufficient representations to classify different emotions occurred in the Arabic tweets, according to [18] finding. [33] have suggested adding semantic features to improve the identification of Twitter sentiments, on the other hand, theproposed model was based on adding the co-occurrence measures, correspond to the contextual information, as additional vectors to support the features' vectors with similar measures for emotions which were expressed using similar features in the classified tweets, which can revel similarity between emotions and therefore help the classifier to distinguish easily between emotions.

The result from this process for both datasets was a collection of features' vectors with their term frequencies. In addition to the similarities between featurising both datasets, few differences in constructing these vectors were encountered.

Following [2]featurisation methodology, labelled features' vectors were generated based on the different terms that were separated by whitespaces in the labelled dataset tweets which were used in the training and the testing sets. On the other hand, features' vectors from the unlabelled set were generated from R normalised rows of C dimensions in the RxC co-occurrence matrix that was built from the unlabelled dataset tweets regardless of the labelled tweets used in the training and learning dataset. These features' vectors consists of frequencies' probabilities between different vocabularies occurred in the unlabelled tweets, where features' vectors from labelled dataset were based on the terms' frequencies of the tweets' terms used during the training and testing processes.

After generating the labelled features' vectors set {kV1,kV2,kV2,kV3….kVz} from the labelled tweets, an equal number of co-occurrence features' vectors set, {rV1, rV2,rV3…..rVz}, were generated. For each labelled feature vector kVi a new feature vector, kVi`, was generated to be used during the classification process which substituted vector kVi. This new generated vector kVi` was resulted from appending a labelled feature vector for each tweet kVi with the unlabelled feature vector rVi that was obtained from combining different rows' co-occurrence vectors kfj of C dimensions of every feature which occurred in kVi, therefor C dimensions were consistent among all the calculated co-occurrence features' vectors, rVi, as well as the co-occurrence features' vectors, kfj, used in generating rVi, on the other hand the number of dimensions for each keyword features' vectors, kVi and therefore the final resulted kVi`, was various for each tweet used in the training and testing process, this variation was based on the number of features generated during the tokenisation process for the classified tweets in the keyword dataset.

Figure 3 1 simplifies the methodology used as well as the featurisation process followed in the proposed model, which is repeated for each tweet used in the training and testing set, in our case this process was repeated 500 times as we used N=500 in each experiment.

Similar to [2] model, the labelled features were generated using n-grams features order, however, co-occurrence features' vectors rVi appended to the labelled features' vectors kVi were based on the occurrence vector for each unigrams features only, kfj in kVi. These features' vectors composed kVi` were used by the SVM classifier which is explained in the following section.

3.5 Classifying Arabic Tweets in the Labelled Dataset

Throughout all the following experiments which the model have performed, SVM classifier was used to classify emotions in different Arabic tweets. In this emotions' classification model, the main goal of the SVM is to discriminate the rule used during the learning process for separating the different emotions accurately into positive and negative set based on the target emotion class with an optimal separating hyper-plane to attain the minimum error rates over the target emotions' classes. Emotions in this model were classified using the distant supervision learning approach, through predicting a predefined emotions classes based on the labels, hashtags and emoticons, used in annotating the labelled dataset, which were then removed from the classified emotional tweets as well as using the co-occurrence information of the unlabelled data in order to help improve the classification performance.

The proposed model has followed [7] and specially [2] classification approach for classifying the six emotions used in Arabic Twitter messages which was based on SVM with the support of both LibSVM and LibLINEAR libraries. Based on [26] proof of the correlation between the features set and the SVM classification accuracy, this has showed the ability of the SVM in handling a large number of features' vectors and dimensional data. Similar to [7], [2], [29] and [34], classification accuracy was based on the K-fold-cross-validation technique, as [2] and [7]the model used 10-fold-cross-validation, in which the classified data set was divided into 10 equal parts, 9 parts out of the total parts were used for the training purposes while the last part was used for the testing purposes. To ensure accurate result all parts need to be tested, through repeating the process 10 different times where the final accuracy result was the average of the 10 repeated processes.

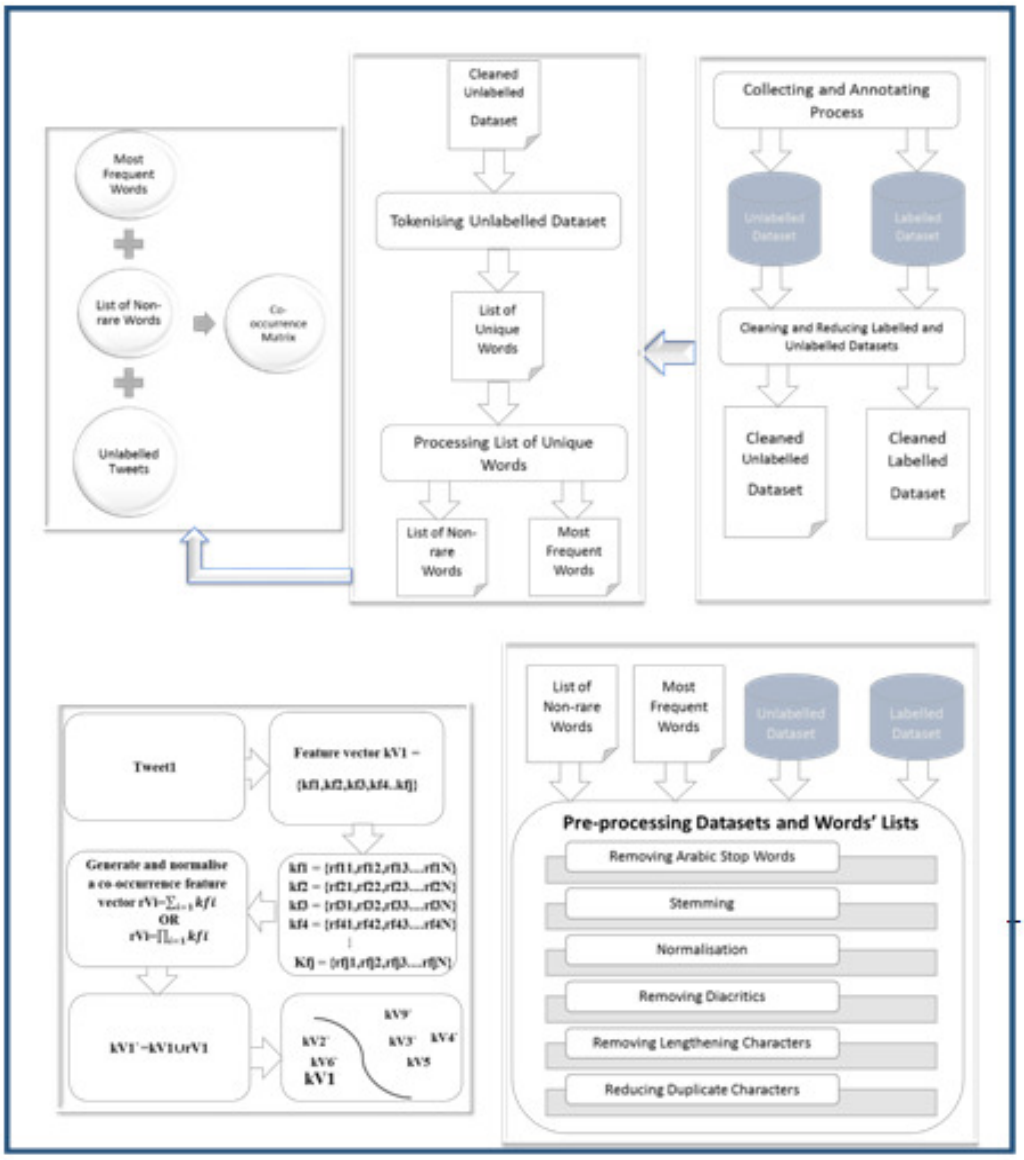

Figure 3 1 Model methodology

4. DATA

Table 4-1 Total unlabeled dataset

UNLABELLED DATASET	
Total	1,368,849 Tweets

Table 0-1 Total labelled dataset

BASE LABELLED DATASET

Happy Hashtag	Sad Hashtag	Anger Hashtag	Fear Hashtag	Surprise Hashtag	Disgust Hashtag
2,758	1,697	*489*	946	578	628
Total			7,096 Tweets		

Happy Emoticon	Sad Emoticon	Anger Emoticon	Fear Emoticon	Surprise Emoticon	Disgust Emoticon
103,741	21,322	*2,381*	4,147	3,068	12,878
Total			147,537 Tweets		

5. EXPERIMENTS AND RESULTS

During this study a number of experiments was carried out to prove our hypothesis through investigating the effect of adding the contextual information of an unlabelled dataset to [2] model, using the proposed methodology from the previous section.

In general two different set of experiments were performed, preliminary experiments as well as a number of underlying experiments based on applying the best techniques investigated during the first set of experiments.

The introductory experiment within the basic set of experiments aimed to test the ability of the built co-occurrence matrix in detecting the similarities between different words and emoticons which exist in the unlabelled dataset. Accordingly, based on the achieved results for the ability of the developed co-occurrence matrix to induce similarities, our methodology was applied to carry the other set of experiments and investigate the effect of the unlabelled dataset on enhancing the classification task, therefore, the following two experiments have explored the effect of both the dimensionality as well as the number of the unlabelled co-occurrence vocabularies used in the co-occurrence matrix on the classifier performance. Following these two experiments, the classifier accuracy was tested with different co-occurrence matrix types binary and frequency co-occurrence matrix. Moreover, as illustrated in the featurisation process, different generated

vectors from the co-occurrence matrix were integrated using different approaches additive and multiplicative approaches, their effect on the classification task was investigated in the fifth experiment of the basic set of experiments. An additional experiment was based on the labelled dataset only without employing our model in including contextual information during the classification process, this experiment was carried out to unify the preprocessing techniques to be applied for the six different emotions with both labels, instead of using different pre-processing schema for each emotion. As a result, the outcomes from these six preliminary experiments were used as the basic features for our developed model to carry out the four additional underlying experiments. After determining the basic model specifications, experiments within the second set were executed to investigate the impact of the previously determined pre-processing techniques set on both datasets for classifying different emotions with the both hashtags and emoticons labels. Furthermore, as Arabic stop words considered the most common occurrence words which do not convey extra meanings, our model has examined their effect by testing the contribution of their contextual information in the classification task before and after removing these words. And finally, the effect of increasing both the labelled dataset as well as the unlabelled dataset was investigated, which was used in the co-occurrence matrix, on the classifier performance to detect different emotions. Some of these experiments are highlighted in this paper.

To guarantee an equal number of positive and negative tested and trained tweets throughout all the different carried experiments and therefore to avoid bias results for popular emotions, the model was tested on a dataset with 500 tweets, N=500, this considered a larger dataset compared to [2] model which used 300 tweets. The selected dataset was divided between positive and negative classes for each experiments equally, N/2 tweets were assigned for each positive and negative class. For the negative class N/2 was divided equally between the other emotions except the target emotion which already has been assigned to N/2 tweets labelled with the target emotion and conventional marker.

5.1 Testing the Effect of the Co-occurrence Matrix Dimensions

The aim of this experiment was to determine the most optimum number of column's and row's dimensions used in the co-occurrence matrix. The decision was based on testing the impact of these dimensions on the accuracy of classifying the different emotions occurred in Arabic tweets of our labelled dataset. In order to confirm the optimum number of dimensions, we have gradually increased the dimensions to measure their effect accurately, where in [32] methodology the co-occurrence dimensions were set directly to the most frequent 10 terms.

chart 5-1 and chart 5-2 illustrate the correlation between the different number of dimensions and the average accuracy for classifying the six emotions with both conventional markers, hashtags and emoticons compared to the accuracy which was resulted from [2] model for classifying these six emotions without employing co-occurrence information from our unlabelled random dataset which was considered as the baseline accuracy throughout the different carried experiments. Regardless to the different number of dimensions tested in this experiment, it was clear from the early results the contribution of including contextual information with different classified features' vectors form the labelled dataset in increasing the accuracy of classifying the different emotions included in the Arabic tweets of our labelled dataset, as increasing these dimensions resulted in adding more contextual information for each feature which occurred in the features' vectors of the different positive and negative tested and trained tweets during the classification process. Although this increase has remained almost steady with the different tested dimensions, it has been proven that including the contextual information of our unlabelled dataset not only

helps in the classification task, but also makes a quite high differences in the classification accuracy, as the average increase was more than +8.5% and +9.53% using different number of the column's and row's dimensions respectively.

Although it is clear the considerable difference between the baselines averages, these differences have been reduced to +0.39% and +0.57% when using our proposed model for including contextual information from our random dataset using different row's and column's dimensions to the classified emotions included in different keyword dataset tweets.

According to the proposed results, 1,000 column dimensions have achieved the highest average accuracy, although 5,000 dimensions were chosen as the optimum column's dimensions from the set of the 100, 500, 1,000, 2,000 and 5,000 tested dimensions to build our co-occurrence matrix. This choice was supported by our belief that the more contextual information we include from the co-occurrence matrix as column's dimensions, the higher accuracy we can achieve. Moreover, as we have built the co-occurrence matrix using 10,000 random tweets only, generalising these 1,000 dimensions as the optimum number of dimensions can considered a risky decision, as the first 10,000 random tweets may fail to represent their strong correlations with the different tested numbers of dimensions. In addition to these reasons, the minor difference, 0.08%, observed from the average accuracies for detecting the different emotions with 1,000 and 5,000 dimensions have provide a further justification to consider 5,000 as the base column's dimensions in our built co-occurrence matrix instead of 1,000 dimensions.

On the other hand, 10,000 vocabularies have been selected for building the constructed cooccurrence matrix. Although 1,000 and 5,000 vocabularies have recorded higher accuracies compared with the classification performance when using 10,000 vocabularies, this selection was based on our aim to maximise the number of co-occurrence feature vectors, denoted as kf1,kf2…kfj in Figure 3-1, for every feature occurred in the positive and negative tweets used during the classification process, hence, this can considered as an inevitable trade-off between the small differences of the achieved accuracies and increasing the number of co-occurrence feature vectors used during the classification and featurisation processes.

On the other hand although 30,000 co-occurrence vocabularies include more vocabularies compared to 10,000 co-occurrence vocabularies, this choice was eliminated as it has been observed from the built co-occurrence matrix a six vocabularies, 0.02%, out of the selected 30,000 did not have any occurrence associations with the previously selected 5,000 dimensions where with the 10,000 vocabularies a 100% associations were found between the vocabularies and the 5,000 co-occurrence matrix dimensions.

As a concluding result, we have observed that by just including a contextual information from 10,000 unlabelled tweets, classifying different emotions has been improved, this has also proved that the information captured from e.g. 100 dimensions and 1,000 co-occurrence vocabularies can be quite enough to produce high performance improvements, which indicates that contextual information gained from adding more hundreds or even thousands dimensions and/or co-occurrence vocabularies does not generate huge differences. Although this cannot be generalised as our contextual information was captured from 10,000 unlabelled random tweets only.

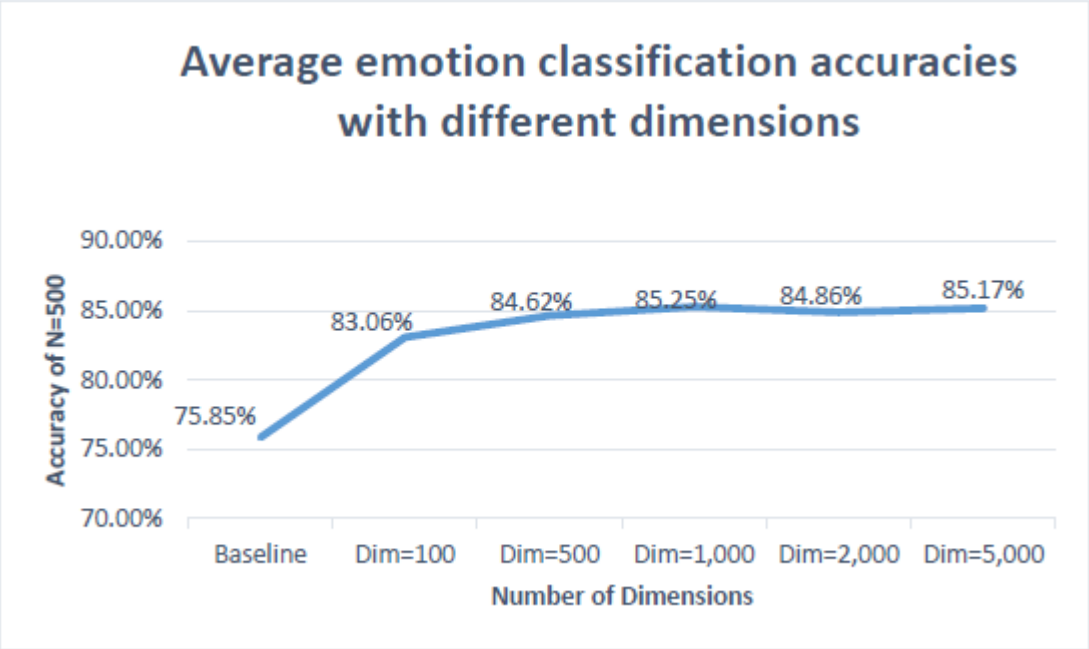

chart 5-1 Co-occurrence matrix effect on the classification accuracy using different column's dimensions

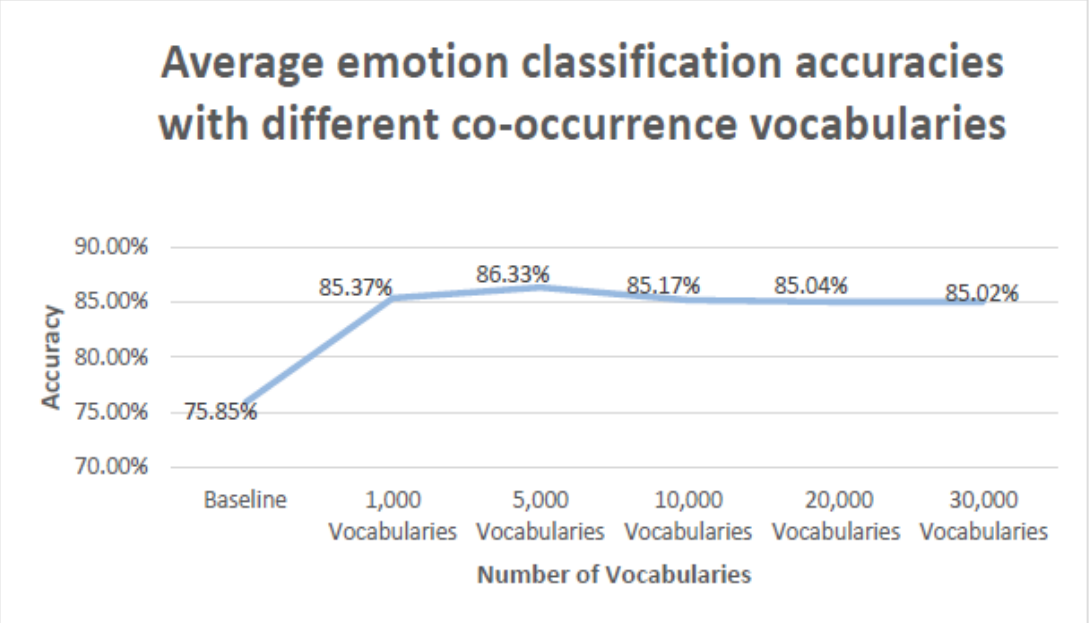

chart 5-2 Co-occurrence matrix effect on the classification accuracy using different row's dimensions

5.2 Removing emotions' labels from the co-occurrence column's dimensions in the co-occurrence matrix

This experiment has answered two questions which were concerned about measuring the popularity of the different labels used in annotating tweets from the labelled dataset as well as the amount of the contextual information these labels can provide.

It has been observed during this experiment that some emoticons have been presented as dimensions in the previously chosen 5,000 dimensions, on the other hand, these dimensions have showed the absence of the different used hashtags for the six target emotions. This can indicate the dominant of the emoticons in Arabic tweets compared to the emotional hashtags which was already proved from the total number of collected tweets using hashtags and emoticons for the six different emotions as shown in Table 4 2Moreover, as emotional hashtags are typed expressions, different users can express the same emotion using different hashtags e.g. (#فرح) and (#فررررح), (#happy) and (#haaaappy) are different hashtags for the same emotion generated from the same adjective (فرح), (happy). Consequently, this has led hashtags to be less probable to appear in our most frequent 5,000 column's dimensions. Therefore this experiment has been performed to the six different emotions labelled with emoticons only.

From the 5,000 column's dimensions only 8 dimensions have been corresponded to different emoticons with a probability of 0.0016% , where 7 of these 8 emoticons [:), :D, (:, :$, :-D, ;), =D, :p],belong to happy class.

Since emoticons are being used essentially as labels in the labelled dataset, using these labels' contextual information during the classification process which explicitly encode the co-occurrence with those same emoticons can generate a bias classification results. Therefore checking that removing the emoticons co-occurrence information doesn't damage the performance is considered an important task which was checked in this experiment through removing the 8 emoticons from the 5,000 dimensions. Consequently, a slight classification decrease has been observed, for classifying emotions with both conventional markers, hashtags and emoticons, from 85.60% in the presence of these 8 emoticons as column's dimensions in the co-occurrence matrix to 85.30% after removing these emoticons. chart 5 3 illustrates the changes between the two approaches compared to the baseline. Although this performance has been dropped compared to the performance when labels information was included in the co-occurrence matrix, the model has continued to improve the classifier performance for all the tested emotions, except for sad and anger, even with the absence of the labels' contextual information from the unlabelled dataset presented as column's dimensions in the co-occurrence matrix.

Therefore, the presented results can prove the strong contribution of the co-occurrence information included during the classification task, in detecting emotions even in the absence of the information associated with the emotions' labels, e.g. emoticons. This strong contribution found is due to the ability for the co-occurrence information to capture related contextual data from the unlabelled tweets which is associated with the different features included in the labelled dataset features' vectors of the classified positive and negative tweets.

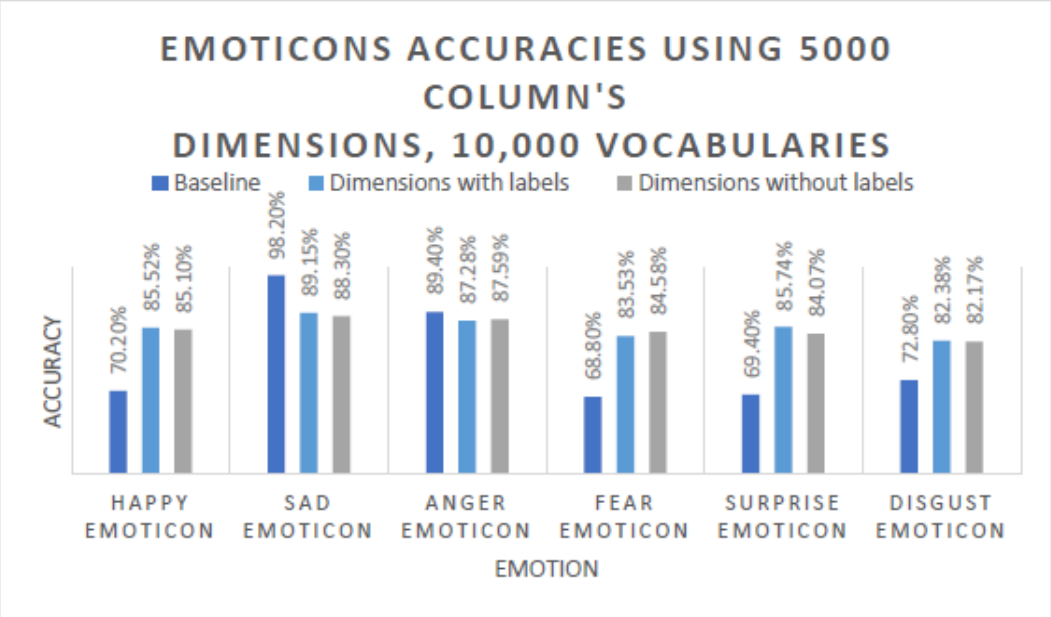

chart 5-3 Emotion classification accuracies using different approaches

5.3 Pre-processing Labelled (Keyword) and Unlabelled (Random) Datasets

This experiment was carried out to investigate the impact of applying the best pre-processing techniques on both datasets, which were inferred from one of the experiments on the labelled dataset. This impact was investigated based on the classification accuracies' changes when co-occurrence information was included during the classification process. The selected set of pre-processing techniques has increased the average performance for classifying hashtags and emoticons from 75.85%, as a baseline average accuracy, to 81.17%. Moreover, this average performance was improved to 85.95% when including pre-processed contextual information of the unlabelled dataset to classify pre-processed tweets from our labelled dataset through applying the best induced techniques.

As illustrated in chart 5 4, applying pre-processing techniques to the different positive and negative labelled dataset and to the co-occurrence information of the unlabelled dataset used in the classification process, have increased the accuracy for classifying hashtags and emoticons with an average increase of +5.02% and + 4.55%. This increase has highlighted the positive impact of compositing the co-occurrence and contextual information of the features with similar contexts under one common feature, on the classification performance, since applying the chosen set of pre-processing techniques have produced common words from different set of words, e.g. in case of applying stemming to the unlabelled dataset, the contextual information of (حزني), (my sadness) and (الحزن) (sadness) combined under (حزن),(sad), similarly the same is applied to the reduction of the repeated characters technique as well as using normalisation technique. Therefore, in the case of pre-processing both datasets, each feature's vector generated from the co-occurrence matrix is a result of combining features' vectors of similar words, which therefore resulting in combining contextual information of similar contexts, referring to the previous example, in case of having (حزني), (my sadness) or (الحزن) (sadness),or (حزن),(sad) as features in

the labelled features' vectors used during the classification, the unlabelled feature vector of the stemmed feature (حزن),(sad) will be used in all the different cases where (حزني), (my sadness) or (الحزن) (sadness),or (حزن),(sad) occurred in the classified tweets. Consequently, this feature vector of the stemmed feature (حزن), (sad) captured the contextual information of all the features originated from this stemmed feature, instead of treating them independently. Although these different pre-processing techniques, except for removing stop words technique, have reduced both the number of row's and column's dimensions of the co-occurrence matrix as they generate number of similar words which were illuminated, their positive impact described previously have prevailed the impact of losing number of repeated dimensions.

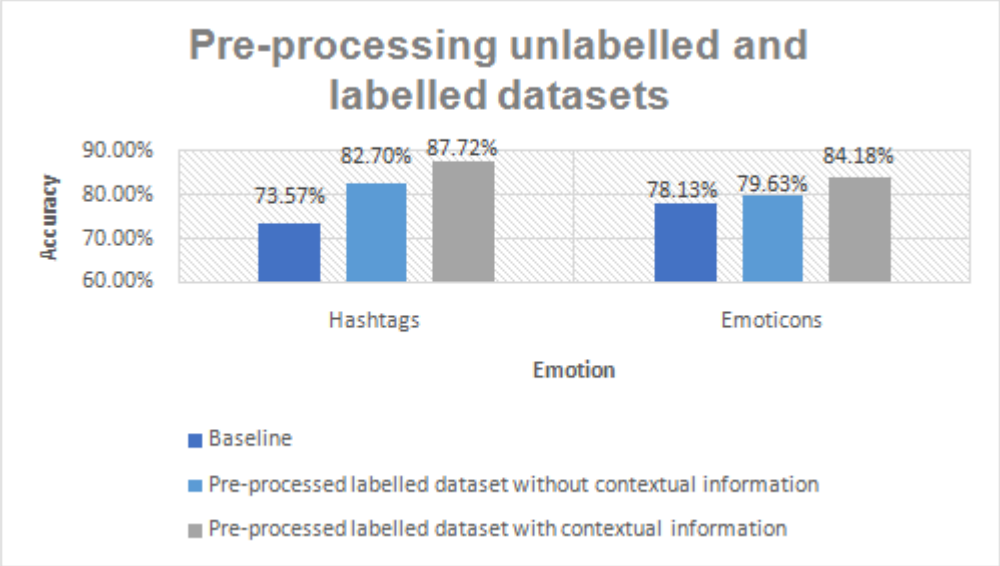

chart 5-4 Classification accuracies before and after pre-process tweets

6. CONCLUSION

During this study a semi-supervised learning approach has been developed to test the effect of the DSM, co-occurrence and contextual information statistics on enhancing the performance of [2] model developed using SVM classifier for the automatic detection of the different emotions' classes found in Arabic tweets collected using hashtags and emoticons for the six target emotions. This has resulted in higher accuracies compared to the standard supervised learning tasks and [2] model, through increasing the amount of the classified texts to include contextual information collected from unlabelled dataset instead of including a large amount of labelled sources, which can be considered inefficient way of increasing the accuracy, due to the time and cost limitations that can be faced in some situations.

Moreover, the model has tested the similarity between different emoticons using co-occurrence information, which proved the sensibility of the captured contextual information. This similarity test has also revealed the sensibility use of the emoticons found in Arabic tweets in relation to the descriptive texts surrounding the different emoticons.

During all the different investigated cases, the model has successfully achieved higher accuracy percentages compared to [2] model when classifying the different hashtags and emoticons for the six emotions, these increases varies based on the different tested factors. Consequently, our experimental results have showed a significant increases in the classification averages for both hashtags and emoticons, which indicate the positive impact of including contextual information during the classification process.

Therefore, based on these achieved results, our model has recorded a higher average accuracy for the six emotions' classes which was more than 86%, compared to most of the highlighted studies which analysed sentiment and emotions in Arabic texts, as illustrated in Table 6 1. On the other hand, a higher accuracy was achieved by [35] which was based on a constructed sentiment dictionary as highlighted previously. Moreover, [3] model have recorded a 95% accuracy which was based on a using a feature selection and extraction algorithm to compute Arabic features' linguistic characteristics, while the developed model did not depend on knowledge sources other than the contextual information of the unlabelled collected tweets.

Table 6-1 The achieved accuracy for some Arabic sentiment and emotion analysis studies

Arabic Sentiment/Emotion Analysis Model	Base/Average Accuracy Achieved
[30]	54% using SVM
[5]	60.5%
[6]	65.87%
[2]	72%

7. FUTURE WORK

As the constructed co-occurrence matrix was built to capture contextual similarities between two different words, contextual information from the unlabelled dataset of bigrams and n-gram features' order included in the labelled set was not included during the classification process. Therefore as a future work, we are planning to extend the co-occurrence matrix to capture their contextual information, as well as to capture the contextual for a larger number of unlabelled tweets to build the co-occurrence matrix. Moreover, we are planning to reflect the effect of different Arabic negation terms in the built co-occurrence matrix. Succeeding [2] model for automating the detection of the different Arabic dialect, we are also planning to extend enhancing the detection of the different emotions on a dialect level using DSM features.

REFERENCES

[1] ASMR, 2014 . Citizen Engagement and Public Services in the Arab World: The Potential of Social Media, Dubai: Mohammed Bin Rashid School of Government.

[2] AlMutawa, B. & Purver, M., 2013. Automatic emotion and dialect detection tool for Arabic language, London : Queen Mary University of London.

[3] Abbasi, A. et al., 2008 . Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Transactions on Information Systems (TOIS), 26(3), pp. 12-46.

[4] AlSubaihin, A. et al., 2011. A proposed sentiment analysis tool for modern Arabic using human-based computing. New York, ACM.

[5] AlSubaihin,A. & AlKhalifa, H., 2014. A System for Sentiment Analysis of Colloquial Arabic Using Human Computation. The Scientific World Journal, 2014(2014).

[6] Abdul-Mageed, M. et al., 2012. SAMAR: a system for subjectivity and sentiment analysis of Arabic social media. PA, USA, Association for Computational Linguistics.

[7] Purver, M. & Battersby, S., 2012. Experimenting with distant supervision for emotion classification. Stroudsburg, Association for Computational Linguistics, pp. 482-491 .

[8] Yuan, Z. & Purver, M., 2012. Predicting Emotion Labels for Chinese Microblog Texts. London, Proceedings of the ECML-PKDD 2012 Workshop on Sentiment Discovery from Affective Data (SDAD).

[9] Plutchik, R., 1980. A general psychoevolutionary theory of emotion. In: Emotion Theory, Research, and Experience. New York: Academic Press.

[10] Sahlgren, M., 2008. The distributional hypothesis. Italian Journal of Linguistics, 20(1), pp. 33-54.

[11] Wiratunga, N. e. a., 2007. Acquiring Word Similarities with Higher Order Association Mining. Berlin, Springer.

[12] Fürstenau , H. & Lapata, M., 2012. Semi-supervised semantic role labeling via structural alignment. Computational Linguistics, 38(1), pp. 135-171 .

[13] Abu-Jbara, A. et al., 2013. Identifying Opinion Subgroups in Arabic Online Discussions. s.l., Proceedings of ACL.

[14] Schutze, H. , 1992. Dimensions of Meaning. Minneapolis, Proceeding of Supercomputing.

[15] Sahlgren, M., 2006. The Word-Space Model , Sweden: University of Stockholm.

[16] Deerwester, S. et al., 1990. Indexing by Latent Semantic Analysis. THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 41(6), pp. 91-407.

[17] Froud, H. et al., 2013. Arabic Text Summarization Based on Latent Semantic Analysis to Enhance Arabic Documents Clustering. International Journal of Data Mining & Knowledge Management Process (IJDKP), 3(1), pp. 79-84.

[18] AlKabi, M. & AlSinjilawi, S., 2007. A Comparative Study of the Efficiency of Different Measures to Classify Arabic Text. University of Sharjah Journal of Pure and Applied Sciences, 4(2), p. 13 – 24.

[19] AlRamahi, M. & Mustafa, S., 2011. N-Gram-Based Techniques for Arabic Text Document Matching; Case Study: Courses Accreditation. ABHATH AL-YARMOUK, 21(1.), pp. 85-105.

[20] Khreisat, L. , 2006. Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. Las Vegas, Proceedings of the 2006 International Conference on Data Mining, DMIN .

[21] ElFishawy, N. et al., 2014. Arabic summarization in Twitter social network. Ain Shams Engineering Journal, 5(2), p. 411–420.

[22] Go, A. et al., 2009. Twitter Sentiment Classification using Distant Supervision. Project Report, Stanford, p. 1–12.

[23] Taboada, M. et al., 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37 (2), pp. 267-307.

[24] Mesleh, A., 2007. Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System. Journal of Computer Science, 3(6), pp. 430-435.

[25] Hmeidi, I et al., 2008. Performance of KNN and SVM classifiers on full word Arabic articles. Advanced Engineering Informatics, 22(1), pp. 106-111 .

[26] Gharib, T. et al., 2009. Arabic Text Classification Using Support Vector Machines. International Journal of Computers and Their Applications, 16(4), pp. 192-199.

[27] Duwairi, R., 2007. Arabic Text Categorization. The International Arab Journal of Information Technology, 4(2), pp. 125-131.

[28] Saad, M. , 2010. The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification, Gaza: The Islamic University.

[29] Rushdi-Saleh, M. et al., 2011. OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology, 62 (10), pp. 2045-2054.

[30] AlKabi, M., AlQudah N. et al., 2013. Arabic / English Sentiment Analysis: An Empirical Study. Irbid, The 4th International Conference on Information and Communication Systems (ICICS).

[31] AlKabi, M. et al., 2014. Opinion Mining and Analysis for Arabic Language. (IJACSA) International Journal of Advanced Computer Science and Applications, 5(5), pp. 181-195.

[32] AlKabi, M., AlBelaili, H. et al., 2013. Keyword Extraction Based on Word Co-Occurrence Statistical Information for Arabic Text. ABHATH AL-YARMOUK, 22(1), pp. 75- 95.

[33] Saif, H. et al., 2012. Semantic sentiment analysis of twitter. Heidelberg , Springer-Verlag, pp. 508-524 .

[34] Chaovalit, P. & Zhou, L., 2005. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. Washington, IEEE Computer Society.

[35] AlKabi, M. et al., 2013. An Opinion Analysis Tool for Colloquial and Standard Arabic. New York, ACM.