ISSN 2278-3091
Volume 7, No.3, May- June 2018
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse04732018.pdf
https://doi.org/10.30534/ijatcse/2018/04732018

An Adaptive Model for Knowledge Mining in Databases "EMO_MINE" for Tweets Emotions Classification

Ibtihal S. Makki, Fahad Alqurashi
King Abdul-Aziz University, Saudi Arabia, ibtihalmakki@hotmail.com
King Abdul-Aziz University, Saudi Arabia, fahad@kau.edu.sa

ABSTRACT

knowledge Mining from large databases has been recognized as a key research topic in database systems and machine learning, many researchers take in their consideration the importance of knowledge extraction from the useful databases. Twitter nowadays becoming a magnificent wide space for getting people's opinions, sentiments and emotions, manufactures aims to knowing their costumers opinions about a specific product, since millions of people connecting through social media pages like twitter for every day and sharing their opinions and emotions towards a product or an event, but there are no enough experiments go further to investigate and characterize the feelings behind tweets. This paper aims to exploring how to facilitate extracting emotions from text tweets by presenting an adaptive model for extracting and classifying emotions in Arabic tweets (EMO_MINE), based on four emotions sad, joy, happy and anger.

The unique value of this model is based on the integration of SQL and machine learning techniques. The experimental results demonstrate how this proposed model for extracting emotions and opinions is useful for knowledge discovery of Arabic tweets using SQL and machine learning algorithms. The results show that our proposed model is improve the classification process in term of accuracy and ROC measurements, the naïve base classifier gives a very satisfy results comparing with others classifiers that's examined in this study.

Key words: Data mining, knowledge discovery, classification, Arabic tweets, Emotions.

1. INTRODUCTION

Data mining is born out of the emergence of new use the large amount of data stored by information of institutions, companies, governments and individuals during the many years. The data becomes the raw material that must be Explode to get a new product, knowledge. This knowledge Becomes a very valuable element for aid in the Decision-making on the area in which they have been collected or extracted the data. While statistics is the first science to consider data as its raw material, to the new needs and the new characteristics of the data (large volume and typology), an important Number of disciplines that begin to integrate what is Known as data mining. Within the data mining there is a branch that is dedicated to the processes of extraction of knowledge whose objective is the knowledge discovery from databases. These processes consist of an iterative sequence of steps or phases. They also present a typology of tasks and techniques to solve it and have measures of Evaluation through the training set. [10]

Emotion classification automizes the process of understanding the deep feeling of humans through their scripts, as demanded in various applications, such as: customer feedback, tourist, e-learning and human computer interaction [5][6][7]. Accordingly, emotion classifying is a very important and demanded, however, this task is nothing but trivial, as will be discussed accordingly. Generally, emotions are viewed as confined levels of sentiments, this is because sentiments categories are more general compared to the way by which the emotions are categorized, as illustrated in Table 1. As such, as the sentiment analysis is known to be hard, emotion extracting and classifying is even more challenging [4]. Besides the fined nature of the emotion categories, the data processed for this task is also overlapped with difficulties. Although document analysis and natural language processing are well established research fields with advanced processing techniques, text analysis for emotion extracting challenged these techniques and demanded for extra processing in order to be able integrated with these techniques. This is because sentiment analysis and emotion analysis depends on analyzing data with special characteristics that is obtained from social networks. In order to analyze such inputs, various challenges have to be faced, such as grammatical mistakes, misspelling, slangs and icons. One of the commonly utilized sources for sentiment analysis and emotion analysis is Tweeter. Approximately, 400 million tweets are posted on tweeter daily, these tweets express emotions that challenge the analysis process [8].

Table 1: Sentiment Classes and the Associated Fined Emotion Categories

Sentiment Class	Emotion Categories
Negative	Anger, Fear, Sadness, Disinterest, Disgust
Positive	Joy, Relief, Excitement, Love
Negative/Positive	Embarrassed, Guilty, Pride, Shame

Existing emotion classification approaches were proposed to process inputs of English language, due to many reasons, most importantly, is that the applications of emotion extraction are mostly targeting English tweets. Yet, there is enormous applications that can benefits from emotion extraction from

Arabic tweets. This is because Arabic is the official language for 58 countries with more than 300 million speakers. Accordingly, there is a need to classify emotions in Arabic language to extends some of the approaches that implemented course task of sentiment analysis and produced positive/negative categorization of Arabic statements. The difficulty raised here is language-based as the Arabic language is highly derivational and inflectional language. Overall, Arabic emotion classification is challenging task, yet it is critical and important to support the development of various fields among Arabic language speakers [5].

The existing approaches for other languages can support Arabic emotion classification, however, the pure machine learning approach that is mostly utilized in this domain has various disadvantages due to its dependency on the size and the quality of the training data and the ignorance of the embedded knowledge revealed by the processed data. Accordingly, a knowledge-mining "EMO_MINE" approach is proposed in this paper. The proposed approach used the commonly utilized TF-IDF calculation to extract features from the entire dataset and then implement a feature selection process based on information gained from each feature in order to eliminate features that do not contribute to the target knowledge. The final step in the proposed approach is using LIBLINEAR SVM and Naïve Bayes (NB) classifiers to classify tweets into four classes, these are: happy, anger, sad and joy. In the implementation of the proposed approach, MySQL is integrated with WEKA data mining tools in order to facilitate feature extraction, selection and mining. MySQL facilitate feature extraction using a set of selection and refinement statements, while WEKA facilitates the process of classification, validation and evaluation.

This paper is organized as the following: A background on the significance of the problem and the utilized processing phases are given in Section 2. Then, the related work is discussed in Section 3. The proposed work and its implementation details are given in Section 4 and Section 5, respectively. Section 6 presents the results of the proposed approach. Finally, the conclusion is given in Section 7.

2. BACKGROUND

The task of emotion extracting and classifying is build based on existing and well-established fields of data mining, natural language processing and document analysis. Accordingly, successful emotion classifying required careful investigation into these fields, processing steps and applications, as will be discussed in this section.

Data mining technology have the benefit of obtaining useful information from numerous databases. Text data mining is the process of "extracting interesting and non-trivial patterns or knowledge from text documents"[9]. Text documents are in semi-structured or unstructured format datasets such as emails, full-text documents, HTML files etc.

Knowledge extraction is mainly related to the discovery process known as Knowledge Discovery in Databases (KDD), which refers to the non-trivial process of discovering knowledge and potentially useful information within the data contained in some information repository [1]. It is not an automatic process, it is an iterative process that exhaustively explores very large volumes of data to determine relationships. It is a process that extracts quality information that can be used to draw conclusions based on relationships or models within the data [2].

Knowledge Discovery from Text (KDT) difficulty lays on extracting explicit and implicit concepts, semantic relations between these concepts, and semantic relations between concepts using techniques of Natural Language Processing (NLP). The main goal is to get insights into large quantities of text data base [10]. When user looks up for something, he will search for it in a traditional way by looking for already known terms and has been written by someone else. The problem lays in the sometimes irrelevant results to the user needs. For that reason, the basic aim for text mining "discover unknown information which is not known and yet not written down"[9].

Knowledge extraction and data mining techniques focus on computer assisted extraction of useful knowledge from data and information. They help to discover and identify hidden patterns (not obvious and sometimes unexpected) in the data that are understandable (Which is especially important in the case of large information funds, where available human resources may be limited). The use of these techniques in multimedia applications is a powerful mechanism to improve understanding and to add value to large repositories of multimedia information. [29]

Data mining is just an essential step which it is objective is discovery of knowledge from a data. This process consists of an iterative sequence of steps or Phases: data preparation, data mining, evaluation, dissemination and Use of models. The extraction of knowledge is an iterative process since the output of some of the phases can return to previous steps and because several iterations are often required to extract knowledge with high quality. It is interactive because the user or an expert in the domain of the problem should assist in data preparation, validation of the extracted knowledge. The following figure illustrates the steps in the traditional KDD process.

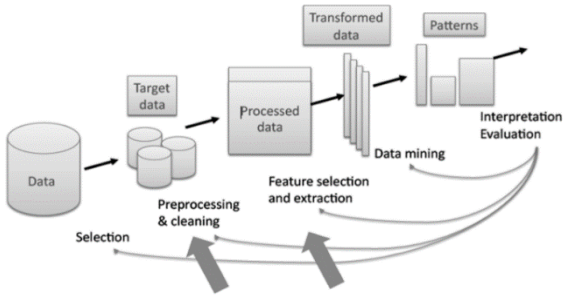

Figure1: Knowledge Discovery from Database process

As shown in the previous figure, the stages of the KDD process are divided into 5 phases and are:

Selection of data, at this stage the sources of data and the type of information to be used are determined. It is the stage where the relevant data for the analysis are extracted from the data source (s) [3].

Preprocessing, this stage consists of the preparation and cleaning of the data extracted from the different data sources in a manageable form, necessary for the later phases. At this stage, various strategies are used to manage missing or blank data, inconsistent data or out of range, obtaining at the end a suitable data structure for later transformation [3].

Transformation, it consists of the preliminary treatment of data, transformation and generation of new variables from the existing ones with an appropriate data structure. Here aggregation or normalization operations are performed, consolidating the data in a necessary way for the next phase [3].

Data Mining, it is the modeling phase itself, where intelligent methods are applied with the aim of extracting previously unknown, valid, new, potentially useful and understandable patterns that are contained or "hidden" in the data [3].

Interpretation and Evaluation, the patterns obtained are identified based on some measurements and an evaluation of the results obtained [3].

Data mining can be applied to any type of information, Being the mining techniques different for each of them. Exist Many data types (integers, real, dates, text strings, Etc.) and from the point of view of the techniques of data mining it would be more usual only to distinguish between two types: numerical (integers or real) and categorical or discrete (they take values in a finite set of categories). Even considering only these two types of data, it should be clarified that not all techniques are capable of working with Both types. These data are contained in what is known as the database, which can be different types depending on the type of information they store. Here we used relational databases which can be described as a collection of data items organized into a set of formally described tables from which data can be accessed or reassembled in many different ways without having to rearrange the tables in the database [19]. the most used today in as a source for data mining techniques. One base of relational data is a collection of relationships (tables) Where each table consists of a set of attributes and can Have a large number of tuples, records or rows. Each tuple Represents an object, which is described through the values of its attributes, and is generally characterized by having a Unique key that identifies it univocally of the rest. One of the main characteristics of relational databases Is the existence of an associated schema, that is, the data Must follow a structure and are, therefore, structured. By means of a query (for example in SQL) we can combine in a single table the information of several tables.

2.1 Emotions

Emotion has various definitions in the literature, some of these are presented here: Emotion is “one aspect of human behavior which plays an important role in human feeling and decision making, thus influencing the way people interact in the society” [5] and “emotions are a mandatory part of human nature that can be considered as hereditary”. [12]. Generally, emotions can be analyzed and categorized using two models, these are the categorical and the dimension, “each type of model helps to convey a unique aspect of human emotion and both of them can provide insight into how emotions are represented and interpreted within the human mind” [11].

Categorical model gained vast interest in the field of machine learning, for its critical applications in various fields, such as: Costumer services, in which emotion classification is used to obtain information about customer satisfaction about services and products. Emotion classification can be integrated with E-Learning in order to create an Intelligent Tutoring System (ITS) that customized the learning process based on the user emotions. In social media, emotion classification is used to customize the recommendations. Most importantly, the field of Human Computer Interaction (HCI), is built based on understand human interaction which could involve emotions [5] [13] [14].

2.1 Emotion Analysis Process Flow

Figure 1 summarizes the main steps involved in the process of Emotion analysis of the text.

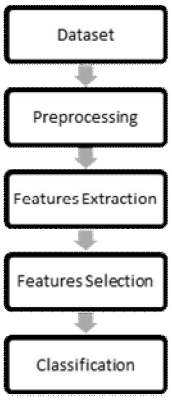

Figure 2: Main Process of Analysis of Feelings.

2.1.1 Preprocessing

The preprocessing step is responsible for cleaning the inputs and transfer the data into a form that can be used successfully by the following processing steps. Generally, the pre-processing involves the following tasks: Filtering, tokenization, stemming, negative removal, stop-word removal and slangs transformation. The filtering and tokenization is responsible for processing the input, remove puncuations, symbols and numbers and produce the contained words. Stemming is an important step in language processing, especially in Arabic language, as it revert words into their roots. Stemming eliminates the variation of the words and

allow for word matching, regardless its derivation from. Finally, slangs transformation convert slang words into words that can be found in the dictionary. Accordingly, the pre-processing step process input with word variations, symbols, numbers and punctuation, eliminates all unutilized parts and convert the rest of the words into abstract form, known as the root, with equivalent dictionary entries. Although these steps are commonly utilized, the implementation of each of them, differ based on the characteristics of the input, the language and the form of the desired output [16] [17][18].

2.1.2 Features Extraction and Selection

As similar to any classification task, the accuracy of the classification outputs depends on the set of features that are extracted from the pre-processed data. In short text and blogs processing, the extracted features can be broadly classified into four categories, these are: syntactic, statistical, semantic and domain-specific features. Using any of these feature extraction, the pre-processed text is converted into a feature vector [19]. One of the commonly and successfully utilized feature is the word significant, which commonly utilized Term Frequency-Inverse Document Frequency (TF-IDF) to calculate the significant of each word according to its frequency in the input and inverse frequency in the corpus/dataset. Several other features are utilized, such as semantic closeness and so on.

Feature extraction, however, must be followed by a feature selection process, that eliminates features with no or bad influence on the classification process and reduce the dimensionality of the feature vector. For the TF-IDF, Mutual Information (MI), Information Gain (IG), and Chi square methods are utilized. Other feature extraction techniques and approaches can be used based on the category of the extracted features [17].

2.1.3 Classification

Given that the input is represented by firmed feature vector, emotion classification can be implemented on these vectors. Existing emotion classification approaches can be categorized into one of the following approaches, machine learning approach and the lexical-based approach [19].

3. RELATED WORK

Various approaches for emotion analysis or similar other problems were proposed in the literature, which can be classified into semantic, combined syntactic & machine learning based and hybrid (combined semantic & machine learning).

In the semantic approach, polarities of words, phrases and sentences are used as features and the process does not require a machine learning algorithm. Researchers in [23] aimd to categorize review into two-classes, these are recommended (thumps up) and not-recommended (thumps down) for four different domains, these are: banks, movies, holiday-destinations and automobiles. Review categorization is implemented by extracting, for the input text, the phrases' polarity, which is a numerical or textual description of words founded in polarity dictionaries. The average polarities of all the phrases are calculated and the input is classified as recommended if the average polarity is above some threshold, otherwise the input is considered as not-recommended. Various other approaches for positive-negative classification with varied input were proposed, such as Liu et al. [33] and Taboada et al. [34]. The problem with these approaches embodied in using two course categories with positive and negative labels. Koppel et al., [25] proved that as a neural class is added to these positive and negative classes in the classification task, more accurate results will be obtained. This suggests that fine classification is not only required as it demanded by the utilized applications, but also to enhance the results.

Although, semantic approaches for other languages provided an acceptable result, the performance of the semantic approach for the Arabic language is not as demanded. In order to assess the efficiency of using a dictionary-based polarities, Rabab'ah et al. [26] evaluates SentiStrength on Arabic datasets of 98,925 instances, which includes, tweets, reviews, comments and posts. The output results have proven to be high, according to some measures, such as precision, however, the accuracy was around 62%, which suggest further research on using SentiStrength with Arabic emotion and sentiment analysis.

In syntactic-based processing, Pang et al. [24], Proposed a model to categorize movies review into two-classes, these are positive and negative. Review categorization is implemented by extracting, for the input text, unigrams, bigrams and other syntactic features, then a number of classification algorithms are utilized, these are SVM, Naive Bayes and Maximum Entropy. Unfortunately, the results obtained from using the developed approach is not as good as it should be. Accordingly, it can be concluded that syntactic features cannot be used solely for emotion analysis, even with a very course task (i.e.: positive and negative discrimination). Using syntactic features, Kiritchenko [35] proposed a model that predicts the polarity of tweets and SMS using the word components. The output, which depends on SVM classified inputs into three categories, positive, negative and neutral. Similarly, Krebs [36] proposed a model that predicts the reaction on Facebook posts using the word components of the post. The reaction is categorized, using Neural Network (NN) into seven emotion classes.

Hybrid approach used semantic polarities with machine learning, in [27] authors proposed a method that used a customized dictionary of positive and negative words in combination with SVM and KNN classifiers. The data which were collected from tweets contained 1000 instances, 500 instances and 500 instances, under sport, social and political topics. The results of the proposed approach were compared to combined syntactic & machine learning and it showed that the proposed approach over performed the syntactic approach.

Similarly, Aldayel and Azmi [28] proposed an approach for analyzing social posts for data that is collected from Saudi Arabia into positive and negative classes.

4. EXPERIMENT

The Experiment used to complete this work was done by using MySQL application which is well known relational database application MySQL is widely used in web applications such as Drupal or phpBB, in platforms (Linux / WindowsApache-MySQL-PHP / Perl / Python), "MySQL is the world's most popular open source database. With its proven performance, reliability and ease-of-use, MySQL has become the leading database choice for web-based applications, used by high profile web properties including Facebook, Twitter, YouTube, Yahoo! and many more. Oracle drives MySQL innovation, delivering new capabilities to power next generation web, cloud, mobile and embedded applications."[20], with the integration with weka software [21], (Waikato Environment for Knowledge Analysis) is an open source program that developed by the University of Waikato, New Zealand, whose objective is to make available, in a simple and transparent way the user, the main machine learning algorithms for use in data mining tasks, automatic sorting, etc. Weka is implemented Completely in the Java programming language, and has a graphical interface and a code interface for execution directly from a development environment. Weka presents tools for pre-processing of data, classification, clustering, association of rules and visualization, and have the feature of being integrated with SQL development tools such as Oracle, MySQL, MS SQL Server.

The experiment was done in a machine with the following hardware configuration: Intel Core i7 CPU, 2.50 GHz and 8 GB RAM. The operating system was windows 10. And the used data set is provided under the *CSIT -2016* [26] after converting it to SQL file manually.

4.1. EMO_MINE Model

Presenting an appropriate model for the opinions, knowledge and emotions extraction from the tweets is the core of success of the whole process. The objective is to extract emotion and knowledge that the user can use by using an integration of SQL and Data Base techniques. This is done by building a model based on the data collected for this purpose. the model is a description of the emotion extraction process that should be done to complete the predictions.

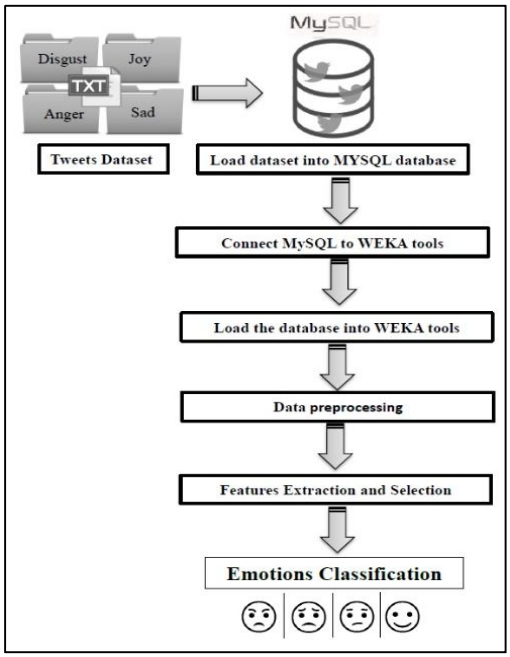

Figure 3: Arabic Emotion Classification (EMO_MINE) Model.

In this section, the proposed model for emotion analysis "EMO_MINE" model will be presented.

While opinion mining can be characterized in terms of three major tasks: a) preprocessing, b) feature extraction and selection, and c) classify the emotion of feeling. And by taking it from the database perspective, to extract emotion from the tweets, EMO_MINE starts by configuring the database and WEKA tools, creating database and its tables, by using create tables queries and loading the tweets database file by using LOAD Query, the classification process

in details as shown in Figure 3. The first step in EMO_MINE is to find an appropriate tweets dataset to analyze its emotions, her we looking for Arabic tweets , and we collecting it as mentioned before in [26]. The second step is to convert the dataset txt file into rff weka file, then converted it into SQL database file, the third step is to create a database into MySQL and create required tables and then load the tweets database file by using LOAD Query, after that the most complicate step is to connect MYSQL to WEKA machine learning tools, next is to load the database file into WEKA explore, that's done by using specific queries to create, show, or even update the tweets database. Next steps is to take advantage of all of the available machine learning algorithms available to classify and cluster the tweets SQL database and those steps are data preprocessing, features extraction, features selection and finally classification. The part of using WEKA tools to complete classification process is discussed in details in A,B,C and D.

A. Dataset Preprocessing

The goal of the pre-processing step is to clean the tweets database and transfer the data into a form that can be used successfully by the following processing steps. As the developed approach is devoted for Arabic tweet, the pre-processing techniques that are utilized to deal with the characteristics of the Arabic language. Overall, the set of the pre-processing steps are as follows:

- ***Data Cleaning***: In this step all non-words components are removed, this includes numbers, symbols, non-Arabic characters and special characters. Besides, stop words are removed from the input as they are not significant to the emotion classification task.
- ***Word Extraction:*** The words components of the text are extracted by a tokenization process using an integration of *TweetNLP* tokenize and TweeQL, "TweeQL can accept arrayor table-valued attributes as arguments. This is required because APIs often allow a variable number of parameters." to deal with twitter, in this work, the tweeQL tokenize returns an array of words that appear in the tweet text.

  For example, the query used in the tokenizer step is,

```
SELECT tweetid, tokenize(text)
FROM sad_tweets;
```

  The tokenize for ("قف على ناصية الحلم و قاتل") = ["قاتل", "قف", "على", "ناصية", "الحلم", "و"].
- ***Word Stemming***: Among various stemming techniques that are presented in the literature, light stemmer [30] is selected to convert the resulted words into their roots. Light stemmer has two significant advantages, these are: 1) Fast, as it is simply implemented by removing prefixes, suffixes from the words without any complex grammatical analysis. 2) Accurate, as its output has proved to be competed with the state-of-the-art and over performs man complex stemmers.

The significance of data cleaning step is two-folds, these are:

- **Reduce the time and space requirements**: As the number of components are decreased, the complexity of the required processing is decreased and the time and memory required to extract and save the feature vector is also decreased.
- **Enhance the accuracy of the classification output**: Having an enormous number of insignificant features in the feature vectors are known to have a bad influence on the classification algorithms. Thus, elimination such bad feature enhances the accuracy of the classification task.

Experimentally, it was found that removing non-words components, over the selected tweet dataset reduced dimension of the feature vector from 7718 to 5494 and removing stop-words further reduce the dimension into 4485. Besides, stemming grouped multiple derivation of a single word into their unique root, which reduce the dimensionality by 30% to 50%.

B. Features Extraction and Selection

After pre-processing, each input in the corpus is represented by a bag of roots, however, each input would be represented using different roots. In order to unify the representation of all the inputs, they are represented using all roots in the corpus. A feature vector is created for each input with a length that is equal to the number of the roots in the whole corpus. The entries of the vector are the TF-IDF for the involved roots, each is represented as a value in the range [0-1]. TF-IDF is calculated as the frequency of the root in the underlying input divided by the frequency of the root in the whole corpus. Accordingly, when the term is appeared frequently in a specific input only, that term will have a high TF-IDF value and considered as significant. On the other hand, a term that appeared in all the input throughout the corpus is considered of low significant and will have low TF-IDF value.

As each input is represented by a complete feature vector, the feature selection process is implemented. The feature selection process that is carried out in the proposed approach depends on the information gain. Accordingly, the information gain of each feature in association with the class labels is calculated. Features with high significance to the classification task are getting higher IG value. As the IG of all features are calculated, these features are ranked, a threshold value is determined and all features with IG greater than the threshold are formed the subset of features that used as input to the classification process.

C. Classification and evaluation

As the inputs are processed, features are extracted and a subset of the most significant features is selected, the actual classification task is conducted. Generally, the accuracy classification task depends on two criteria, these are the features, which were discussed previously and the classification algorithm. In the proposed approach, three classification algorithms are utilized, these are:

- **Support Vector Machine (SVM)**: SVM classifier is a well-known algorithm for its accurate output in complex classification and regression task, accordingly it has been selected to be implemented with the proposed approach. However, SVM suffers from its complexity with multi-class classification problem with sparse input. The implementation of this algorithm is achieved using LIBLINEAR SVM [31].
- **Naïve Bayes (NB) [1]**: A probabilistic classifier that is built based on the independence assumption among the involved feature set. The advantages of NP are the accuracy, simplicity and its low computational complexity even with multi-class classification problem with sparse input, accordingly it has been selected to be implemented with the proposed approach.

- **J48**: A tree based classification algorithm that is known to give an accurate output and compete with the previously discussed algorithms.

The experiments are formed by dividing the entire dataset into training and testing sets in various ways, these are:

- Splitting 75%-25%: Dividing the dataset into two subsets, such as 75% of the dataset is used for training and the rest for testing. The accuracy is evaluated based on the correctly classified instances in the testing set (25% of the entire dataset).
- Splitting 65%-35%: Dividing the dataset into two subsets, such as 65% of the dataset is used for training and the rest for testing. The accuracy is evaluated based on the correctly classified instances in the testing set (35% of the entire dataset).
- Splitting 60%-40%: Dividing the dataset into two subsets, such as 60% of the dataset is used for training and the rest for testing. The accuracy is evaluated based on the correctly classified instances in the testing set (40% of the entire dataset).
- Splitting 55%-45%: Dividing the dataset into two subsets, such as 60% of the dataset is used for training and the rest for testing. The accuracy is evaluated based on the correctly classified instances in the testing set (40% of the entire dataset).

- 10-fold cross validation: Dividing the dataset into two subsets, such as 90% of the dataset is used for training and the rest for testing in the first fold and save the results for later usage. Then divide the set again with the same percentage, but using different data for the testing. As the 10 runs are implemented, the entire dataset would be tested in different folds. Accordingly, the accuracy is evaluated based on the correctly classified instances in the entire dataset.

5. RESULTS and DISCUSSION

As previously mentioned, two experiments were done to complete the classification process in this work, Percentage split and k-fold cross-validation, in the Percentage split experiments four varied training/testing datasets were applied: (55%–45%), (60%-40%), (65%-35%) and (75%–25%) to decide which splitting ratio is the best. The experiment shows that, the results get better when using the **(75%–25%)** split ratio which gives high accuracy as shown in table 2.

In the k-fold cross validation, the second test method. authors use 10-fold cross validation, in this mode, dataset is separated into 10 sections (folds), hold out each part thus and take the normal of all outcome. Each part is utilized once to test and 9 times for preparing. Table 6 indicates test-mode (10-Fold Cross-Validation). Table 8 indicates test-mode (10-Fold Cross-Validation). This paper center around the 10-Fold Cross-Validation test technique, since it gives the most greatest accuracy.

Table 1 and table 2 shows the used performance metrics to evaluate EMO_MINE model. Which are Kappa statistic, Mean Absolute Error (MAE) and a Receiver Operating Characteristic (ROC) area which are used to measure the accuracy.

Classifier algorithm tools	Correctly classified instances (%)	Kappa statistic	MAE (%)	(Weighted Avg.) ROC
LIBLINEAR SVM	**74.642**	0.634	0.125	**0.917**
Naïve Bayes	**73.134**	0.621	0.164	**0.933**
J48	48.763	0.208	0.292	0.771

Table 2: split ratio test mode (split 75% train, 25% test).

Table 3: cross-validation test mode (10-fold cross-validation).

Classifier algorithm tools	Correctly classified instances (%)	Kappa statistic	MAE (%)	(Weighted Avg.) ROC
LIBLINEAR SVM	**87.314**	0.744	0.154	**0.984**
Naïve Bayes	**85.657**	0.733	0.163	**1**
J48	50.348	0.286	0.334	0.724

The kappa **statistic** measurement is a metric that as often as possible used to test inter-rater reliability as a connection measurement to measure the accuracy of the overall model, kappa can extend from -1 to +1. where 0 represents to the normal expected value, 1 represents to perfect value. The **mean absolute error (MAE)** is a metric used to gauge how close conjectures or expectations are the possible results, here, the smaller result value is better. (ROC curve) is a graphical plot that clear up the execution of a paired classifier framework as its segregation of various thresholds. The Area under the ROC bend is a powerful method to measure precision and accuracy, an area of 1 represents a perfect test result. [32]

As shown in in table 2 according to the10-fold cross validation test method, It can be clearly seen that,:

- The model reached higher accuracy results while using the LIBLINEAR SVM and Naïve Bayes classifier algorithm, which is considered as a good accuracy level for the model.
- The model achieved higher measurements while using LIBLINEAR SVM classifier algorithm, than Naïve Bayes and J48 classifiers with 87.3% accuracy and 0.744 Kappa statistics.
- The model obtained best result in ROC while using Naïve Bayes classifier than LIBLINEAR SVM classifier with 1 Weighted Average. Since Naïve Bayes classifier reached highest area under the ROC value, which is 1, and regarding this situation, this classifier will consider to be the best classifier for use in this model.

As shown in Figure 4 according to the10-fold cross validation test method and regarding the accuracy of all classifiers, it can be clearly seen that

- The model reached 87.314 % of accuracy while using the LIBLINEAR SVM classifier algorithm, which is considered as the best accuracy on the overall classifiers used in this model.
- When using Naïve Bayes classifier algorithm, the model obtained 85.657% of accuracy, which considered as a good accuracy level for the model.
- The model reached 50.348 % of accuracy while using the J48 classifier algorithm, which **can't** be an acceptable accuracy level for the model.

Figure 4 shows the accuracy measures rate of all classifiers in the10-fold cross validation test method,

Figure 5. shows different Measurements Rate and figure 6 shows the ROC curve for the three classifier algorithms

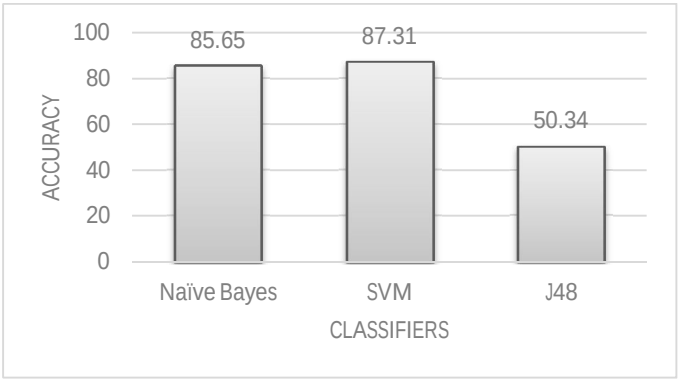

Figure 4: Correctly Classified Instances.

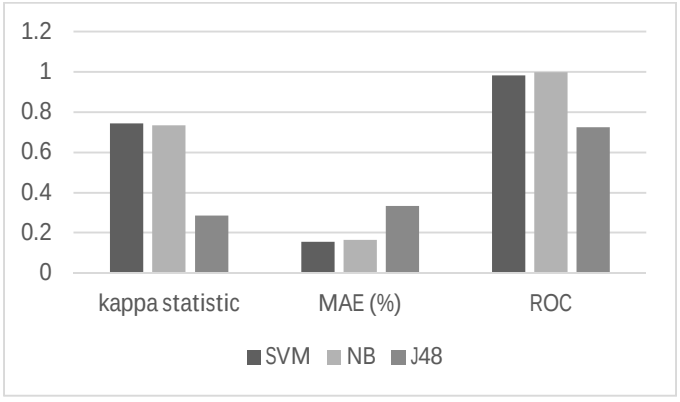

Figure 5: Rate of different Measurements

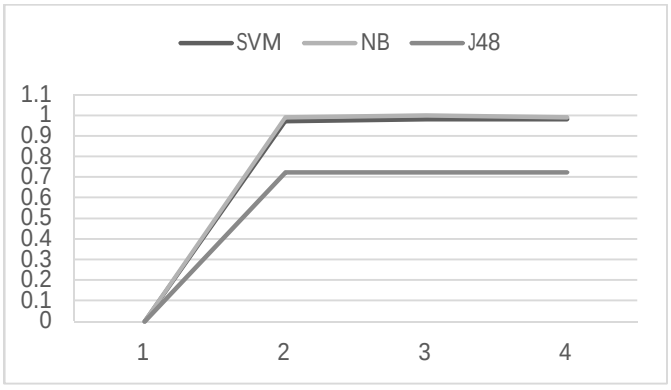

Figure 6: ROC curve for the three classifier algorithms

6. CONCLUSION and FUTURE WORK

The adaptive "EMO_MINE" model is proposed in this work, based on using SQL relational databases to import the tweets data and process it by using SQL Queries and WEKA platform for machine learning techniques, Analyze and Classify emotions in Arabic tweets that's selected randomly to four classes of emotions, sad, happy, joy and anger.

The use of the integration of SQL with machine learning algorithms and the use of SQL Queries improve the classification process and overcome some limitation of the traditional methods in dealing with Arabic language Due to the complexity of it, authors develop EMO_MINE model using common classification algorithms, to complete the classification process, where Naïve Bayes classifier gives the highest ROC results and a satisfy accuracy. EMO_MINE model enables the classification to be done with a high accuracy rate which is 85%. The results of this work demonstrate that the proposed model give a satisfy results by classifying an Arabic tweets dataset into four main emotions which are sad, joy, happy and anger.

REFERNCES

1. A. Shukla and S. Shukla, "A Survey on Sentiment Classification and Analysis using Data Mining," *Int. J. Adv. Res. Comput. Sci.*, vol. 6, no. 7, pp. 20–24, 2015.
2. A. Assiri, A. Emam, and H. Aldossari, "Arabic Sentiment Analysis: A Survey," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 12, 2015.
3. M. Hasan, E. Rundensteiner, and E. Agu, "EMOTEX: Detecting Emotions in Twitter Messages," *ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conf.*, pp. 27–31, 2014.
4. M. A. Al - and M. Al - Ayyoub Jordan, "A Lexicon - Based Approach for Emotion Analysis of Arabic Social Media Content," no. June, 2016.
5. M. C. Jain and V. Y. Kulkarni, "TexEmo: Conveying Emotion from Text-The Study," *Int. J. Comput. Appl.*, vol. 86, no. 4, pp. 975–8887, 2014.

6. R. Kumari and M. Sasane, "Emotion analysis using text mining on social networks," *Int. J. Innov. Res. Technol.*, vol. 2, no. 1, pp. 2349–6002, 2015.
7. N. M. Shelke Assistant Professor and P. Indira Gandhi, "Approaches of Emotion Detection from Text," *ISSN*, vol. 2, no. 2.
8. L. Wikarsa and S. N. Thahir, "A text mining application of emotion classifications of Twitter's users using Na??ve Bayes method," in *Proceeding of 2015 1st International Conference on Wireless and Telematics, ICWT 2015*, 2016.
9. S. Vijay Gaikwad, P. D. Y Patil, and P. Patil, "Text Mining Methods and Techniques," *Int. J. Comput. Appl.*, vol. 85, no. 17, pp. 975–8887, 2014.
10. V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *J. Emerg. Technol. Web Intell.*, vol. 1, no. 1, pp. 60–76, 2009. https://doi.org/10.4304/jetwi.1.1.60-76
11. S. Mac Kim, "Recognising Emotions and Sentiments in Text," no. April, p. 128, 2011.
12. S. Dhawan, K. Singh, and D. Sehrawat, "Emotion Mining Techniques in Social Networking Sites," *Int. J. Inf. Comput. Technol.*, vol. 4, no. 12, pp. 1145–1153, 2014.
13. A. G. Shahraki, "Emotion Mining from Text," 2015.
14. N. M. Shelke, "Approaches of Emotion Detection from Text," *Int. J. Comput. Sci. Inf. Technol. Res.*, vol. 2, no. 2, pp. 123–128, 2014.
15. O. Appel, F. Chiclana, and J. Carter, "Main Concepts, State of the Art and Future Research Questions in Sentiment Analysis," *Acta Polytech. Hungarica*, vol. 12, no. 3, pp. 2015–87.
16. A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining."
17. V. B. Vaghela, B. M. Jadav, and M. E. Scholar, "Analysis of Various Sentiment Classification Techniques," *Int. J. Comput. Appl.*, vol. 140, no. 3, pp. 975–8887, 2016.
18. A. Giachanou and F. Crestani, "28 Like It or Not: A Survey of Twitter Sentiment Analysis Methods," *Article*, vol. 49, no. 28, 2016.
19. S. Chen and W. Pedrycz, *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*, vol. 639. 2016.
20. MySQL database management system development tool". [Online]. Available: https://www.mysql.com/. [Accessed: 24- May- 2016].
21. WEKA machine learning tool". [Online]. Available: *https://weka.wikispaces.com/*. [Accessed: 24- May- 2016]
22. A. Mudinas, D. Zhang, and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," *Proc. First Int. Work. Issues Sentim. Discov. Opin. Min. - WISDOM '12*, pp. 1–8, 2012.
23. P. D. Turney, "Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews," *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 417–424, 2002.
24. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," *Empir. Methods Nat. Lang. Process.*, vol. 10, no. July, pp. 79–86, 2002. https://doi.org/10.3115/1118693.1118704
25. M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," *Comput. Intell.*, vol. 22, no. 2, pp. 100–109, 2006. https://doi.org/10.1111/j.1467-8640.2006.00276.x
26. A. M. Rabab'Ah, M. Al-Ayyoub, Y. Jararweh, and M. N. Al-Kabi, "Evaluating SentiStrength for Arabic Sentiment Analysis," *Proc. - CSIT 2016 2016 7th Int. Conf. Comput. Sci. Inf. Technol.*, 2016.
27. S. O. Al-humoud, "Arabic Sentiment Analysis using WEKA a Hybrid Learning Approach Arabic Sentiment Analysis using WEKA a Hybrid Learning," no. November, 2015.
28. H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis – a hybrid scheme," *J. Inf. Sci.*, vol. 42, no. 6, pp. 782–797, 2016. https://doi.org/10.1177/0165551515610513
29. Rajput, Anil, et al. "J48 and JRIP rules for e-governance data." International Journal of Computer Science and Security (IJCSS) 5.2 (2011): 201.
30. L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light Stemming for Arabic Information Retrieval."
31. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," J. Mach. Learn. Res., vol. 9, pp. 1871–1874, 2008.
32. Abdullah, M., AlMasawa, M., Makki, I., Alsolmi, M., & Mahrous, S. (2018). Emotions extraction from Arabic tweets. International Journal of Computers and Applications, 1-15. https://doi.org/10.1080/1206212X.2018.1482395
33. Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342-351). ACM https://doi.org/10.1145/1060745.1060797
34. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307 https://doi.org/10.1162/COLI_a_00049
35. Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research, 50, 723-762.
36. Krebs, F., Lubascher, B., Moers, T., Schaap, P., & Spanakis, G. (2017). Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. arXiv preprint arXiv:1712.03249.