An ML-Based Classification Scheme for Analyzing the Social Network Reviews of Yemeni People

Emran Al-Buraihy
Faculty of information Technology, Beijing University of Technology, China
emran.thabet3003@gmail.com

Wang Dan*
Faculty of information Technology, Beijing University of Technology, China
wangdan@bjut.edu.cn

Rafi Ullah Khan
Institute of Computer Science and Information Technology, The University of Agriculture Peshawar, Pakistan
rafiyz@aup.edu.pk

Mohib Ullah
Institute of Computer Science and Information Technology, The University of Agriculture Peshawar, Pakistan
mohibullah@aup.edu.pk

Abstract: *The social network allows individuals to create public and semi-public web-based profiles to communicate with other users in the network and online interaction sources. Social media sites such as Facebook, Twitter, etc., are prime examples of the social network, which enable people to express their ideas, suggestions, views, and opinions about a particular product, service, political entity, and affairs. This research introduces a Machine Learning-based (ML-based) classification scheme for analyzing the social network reviews of Yemeni people using data mining techniques. A constructed dataset consisting of 2000 MSA and Yemeni dialects records used for training and testing purposes along with a test dataset consisting of 300 Modern Standard Arabic (MSA) and Yemeni dialects records used to demonstrate the capacity of our scheme. Four supervised machine learning algorithms were applied and a comparison was made of performance algorithms based on Accuracy, Recall, Precision and F-measure. The results show that the Support Vector Machine algorithm outperformed the others in terms of Accuracy on both training and testing datasets with 90.65% and 90.00, respectively. It is further noted that the accuracy of the selected algorithms was influenced by noisy and sarcastic opinions.*

Keywords: *Social network, sentiment analysis, Arabic sentiment analysis, MSA, data mining, supervised machine learning.*

Received March 18, 2020; accepted October 31, 2021
https://doi.org/10.34028/iajit/19/6/8

1. Introduction

The development of science, economy, information technology, and communications has led to an increase for data in recent times. These vast amounts of data are no longer able to deal with traditional methods of analysis such as statistical. Data mining has emerged since the 1980s and proved to be one of the most successful solutions for large amounts of data analysis, by converting them from mere the information accumulated to data that can be used to take advantage and utilized thereafter [16]. Such enormous amounts of data can be found in social networks. Social network allows individuals to create public and semi-public web-based profiles within a domain in order that they communicate with other users in the network and online interactions sources, sharing of contents, evaluations, approaches, feelings, opinions, and emotion expressions contained in the text, comments, blogs, news, comments, feedback or such other documents [29]. Social media are a prime example of the social network. The people express their ideas, suggestions, views, and opinions about a particular product, service, political entity and current affairs [11]. The rise of microblogging and social media has led to a rise in the prices of personal review ratings: from critical articles, estimates, nominations and other online expression forms. For computer scientists, this fast-growing heap of information opens a door to the massive awareness of social network users [6].

Opinion mining is one of the major topics related to data mining field, which aims to extract and analyse the feedback of people. The main issue to take into account is to find product feature and comments of analysis could be positive or negative. Generally, people determine predefined terms to express them as positive or negative comments [28]. Sentiment analysis is also known as opinion mining is a method of analysing opinions of people, feelings, assessments, viewpoints, and emotions of the written language. For Natural Language Processing (NLP), Sentiment analysis is considered as one of the most active research areas also involved in many fields such as management sciences, political science, economics, and social sciences and has been extensively studied in data mining and its branches. It primarily focuses on opinions that express whether sentiments positive or negative. With the importance of knowing opinions and feedbacks of the people, sentiment analysis systems have been using almost in every business and

social domain to assist in decision-making [19]. Classification of reviews has been studied deeply in languages like English, Chinese, and Spanish. However, little research studies in this area have been conducted in other languages such as Urdu, Italian, and Arabic [11]. Even most of the conducted research studies in the Arabic language mainly focused on formal Arabic language, and a few of them performed on a local dialect such as Egyptian [4] and Jordanian [4, 7]. Arabic is a Semitic language spoken by more than 330 million people as an indigenous language. Arabic is a highly organized and derivative language, in which morphology has a very important role [27]. A huge number of social media users in the Arab world are expressing their opinions on a variety of domains such as product reviews, political events, sports, etc. Consequently, we need to pay more attention to the analysis of reviews in the Arabic language, especially in Yemeni dialects. In this research, an ML-based classification scheme is introduced for analyzing the social network reviews of Yemeni people. A large Arabic dataset has been constructed, annotated manually and then used for the training and testing of the classification scheme. The developed dataset contains only reviews in Modern Standard Arabic (MSA) and Yemeni dialects, which were extracted from Twitter and Facebook. Then, four supervised machine learning algorithms were applied to the dataset. Besides, a comparison of the performance of the classifiers was considered.

The main contributions of this paper can be summarized as follows:

- A training dataset composed of 2000 MSA and Yemeni dialects unique records used for training and testing purposes has been constructed and processed comprehensively, and automatically annotated.
- A testing dataset composed of 300 MSA and Yemeni dialects unique records used to demonstrate the capability of our scheme has been contracted.
- To train the machine using training dataset and build Decision Tree, Naïve Bayesian, Support Vector Machine (SVM) and K-Nearest Neighbours (K-NN) classification models for constructed datasets.
- To compare the performance of the above classification models based on Accuracy, Recall, Precision and F-measure.

2. Related Work

Several studies have been carried out in the field of sentiment analysis. Interested approaches and different systems have been suggested and developed by the researchers to work around this problem. In an esoteric way, most of these advanced systems are intended for English and cannot be applied to other languages perfectly. In this section, we will discuss several major papers in this area and try to point out the useful works of the Arabic language.

There are three learning approaches for sentiment analysis which are supervised, unsupervised, and semi-supervised. The main difference between supervised and the other two is in building the training dataset. Supervised learning involves some requirements such as Pre-processing, Normalization, while the others do not involve [8]. Abdulla *et al*. [1] presented a hybrid-based approach to sentiment analysis for the Arabic language. A dataset was created, consisting of 2,000 Arabic tweets, and a lexicon created, consisting of 3479 Arabic emotional words. A tool was designed to find the text sentiment orientation by calculating the total weights of the entire input text. SVM, Decision Tree (DT), Naïve Bayes (NB), and KNN algorithms were applied. The experiments showed that SVM outperformed the others. Duwairi and Qarqaz [13] developed a framework, which has a set of dictionaries such as Jordanian dialect, MSA, and Arabizi to deal with Arabic sentiment analysis. Three classifiers were used particularly, SVM, KNN, and NB. They had to use stratified shuffling to avoid memory issues in Rapidminer. The results demonstrated that all the classifiers had good accuracy when stopwords filter and stemming techniques not used. Duwairi *et al*. [12] proposed a classification model by using a supervised machine learning approach to deal with Arabic reviews for sentiment analysis. A dataset was developed, made up of 2591 tweets/comments. SVM, KNN, and NB were applied. The results show that KNN (K=10) gave the highest recall, while SVM gave the highest precision. Mahyoub *et al*. [21] used semi-supervised learning to circulate the scores Arabic Sentiment Lexicon, which is available in Arabic WordNet made up of more than 800 positive, more than 600 negative and more than 6000 natural words. The Arabic WordNet and the sentiment orientation were applied to extend seed lists in the expansion algorithm, while the lexicon was evaluated by the task-based evaluation method. Two Arabic corpora were combined and represented the corpus by using the Vector Space Model algorithm. The classification model was built by the Rapidminer, two machine learning classifiers: (SVM and NB), and applying the TF-IDF algorithm to the classifiers. The experiments showed the superiority of NB in the term of accuracy. Parveen *et al*. [25] proved the importance of addressing the sarcastic tweets to enhance the accuracy. Two datasets were created, one for training consists of 20,000 tweets, and the second for testing consists of1200 tweets. Three classifiers were used particularly, NB, SVM, and ME for performing a classification. The results show a significant improvement after considering sarcasm, and the proposed method outperformed the baseline one in both cases before and after taking sarcasm, while SVM was better than NB and ME in the term of accuracy.

The Alhumoud *et al*. [8] designed a sentiment analyzer to analyze a dataset, which is made up of 3690 sentimental words. Supervised and hybrid approaches were applied, and two algorithms utilized to build the model SVM and KNN. In training the classifier, converting algorithm had to be used to match the attributes. The results showed that hybrid learning outperformed supervised in the term of accuracy. Elnagar [15] investigated on sentiment analysis for Arabic reviews using a supervised classifier model on the Large-Scale Arabic Book Reviews (LABR) dataset books' reviews, which consists of 63257 reviews. SVM and Linear Regression (LR) classifiers were used. IPython Notebook with GraphLab Create were selected for the experimentation setup. The results show slightly superior for the baseline in imbalanced dataset whereas outright superiority for the balanced dataset in term of accuracy. Alayba *et al*. [3] established an Arabic language dataset opinion on health services consists of 2026 Arabic tweets. Deep neural networks algorithm was applied alongside with several Machine Learning algorithms such as SVM, NB and, LR. SVM classifier using Linear Support Vector Classification and Stochastic Gradient Descent obtains the best accuracy. Itani *et al*. [17] proposed a corpus for Sentiment Analysis of Arabic language. Two built corpora consist of 2000 posts. The collected posts of news and art domains consist of 12053 and 8423 words, respectively. Besides that, building a lexicon by extracting the words and phrases from each post consists of 2509 words. The classifier performance was measured by comparing its results along with the annotation, which was done by human taggers and showed promising results. Al-Sukkar *et al*. [9] applied supervised machine learning techniques, namely SVM, NB, and DT on sentiment analysis for Arabic tweets. The results showed a preference for SVM and noting that stemming technique was beneficial to all techniques. Al-Azani and El-Alfy [5] investigated some issues in imbalanced data of Arabic sentiment analysis. A dataset that was built consists of 1798 tweets. The problem of its imbalanced was handled by using Synthetic Minority Over-sampling Technique (SMOTE) which computes the difference between a considering sample with its nearest neighbour, while Word2vec tool was used to compute word vector. The computational models were generated by evaluating many ensemble learning techniques, while other models were generated by using single classifiers, such as K-KNN, SVM, LR, Stochastic Gradient Descent (SGD), Gaussian NB and decision trees with and without SMOTE. The results showed an improvement when SMOTE and the stacking ensemble learning method for F1-score overall average were applied. Mukhtar and Khan [23] proposed a classification model using a supervised machine learning approach to deal with sentiment analysis in the Urdu language. A dataset that was built consists of 6025 sentences. Three algorithms were utilized to build the model, SVM, DT, and KNN using Weka software. The results showed a slight superiority of the KNN algorithm after applying further analyses. Mostafa [22] proposed sentiment analysis approach based on lexicon-based and supervised classification. Three automatic lexicons were developed for storing purposes, MSA, Arabic colloquialism, and negation terms. The experimental results were performed by using supervised machine learning algorithms SVM, K-KNN, and NB. The results were promised in comparison with three recent research papers. El-Masri *et al*. [14] developed a tool that has different techniques for Arabic sentiment analysis. The tool has a set of parameters that allows users to select a specific topic and adjust the desired requirements. It was trained with 8000 of randomly selected Arabic tweets. SVM and NB machine learning algorithms and Lexicon-based approach were considered. The experiments showed that the lexicon-based approach produced good results, despite the small size of the lexicon, while in the Machine Learning approach, NB was better in the accuracy of the polarity prediction topic. Mustafa *et al*. [24] proposed a new hybrid approach for Arabic sentiment analysis. A publicly available corpus is made up of 2967 tweets. Besides, expanding a publicly available lexicon to contain 7358 words and 1527 idioms (Look-Up table), and constructing a lexicon consists of 552 roots. The experimental results showed a clear superiority of the lexicon-based in comparison with the other two different stemming techniques, and reached its higher results when it was used with SVM and NB algorithms.

After studying the literature review, it was clear that varieties of works were carried out on different approaches such as supervised (Corpus-based), unsupervised (Lexicon-based), and semi-supervised (Hybrid) machine learning. Furthermore, it was also observed that the most supervised machine learning algorithms commonly used were SVM, NB, KNN and DT, respectively. In each case of a research study, accuracy, precision, and recall were the performance metrics.

3. The Proposed Scheme

Supervised machine learning was applied to a developed and classified MSA and Yemeni dialects dataset using RapidMiner. The developed dataset and testing dataset are available upon contacting the authors. The proposed method consists of five steps, as shown in Figure 1. The data was collected from Twitter and Facebook involved in the political domain. For the preprocessing purpose, many techniques were used to prepare the dataset for classification. Before the classification process, important features were extracted. Then, clarifying how the classification model was built. In the last step, the results were

analysed based on some performance parameters.

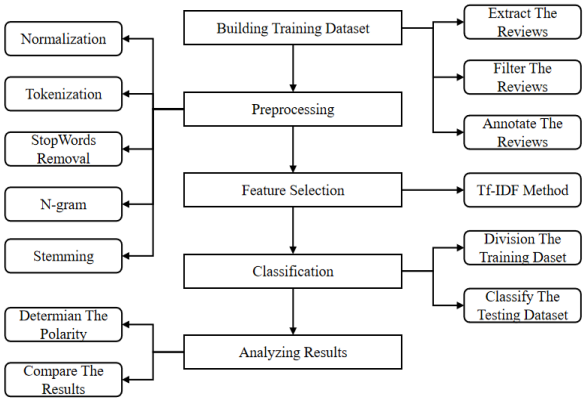


Figure 1. Flow diagram for the proposed scheme.

Algorithm 1: The proposed Scheme

Data: Sort the input Terms T_n
Data: Get all the reviews r_n
Data: Prepare the dataset DS
$R_{set} \leftarrow reviewsT_i$
forall $r_n \in R_{set}$ *do*
  if $r_i \in T_i$ *then*
    $DS \leftarrow r_i$
  End
end
forall $r_n \in DS$ *do*
  $x \leftarrow preprocessing\ r_i$
    $y \leftarrow feature\_selection\ r_i$
end
forall $r_n \in DS$ *do*
  $G_{model_{x,y}} \leftarrow classification\_algorithms(r_i)$
end
Return (output)

In the first step, the user sets some sort of terms/topics (Tn) which the reviews (*rn*) are extracted based on them. Second, the *Rset* file contains all the extracted reviews (*rn*). Then, all the reviews (*rn*) in the *Rset* file those which actually related to a particular term (*Ti*), saved in a Dataset (DS) i.e, the DS only contains the reviews (*rn*) which in fact belong to the entered term. Afterward, all the reviews (*rn*) in the DS are pure and ready to be forwarded to build the model. Hence, all the reviews (*rn*) prepared to go through multiple pre-processing steps. Furthermore, important features are extracted and then used as inputs for the classifiers. Then, the models are built based on the selected classifiers. Finally, every review (*rı*) in the DS given a polarity, and the results were displayed.

3.1. Building Training Dataset

The training dataset was built by collecting tweets and Facebook comments consists of reviews divided into 1307 and 693 of negative and positive reviews, respectively. All the tweets and comments written only in MSA and Yemeni dialects, and involved in the political domain. In this study, aspect level performed on some of the very long reviews which contain more than a feeling expression. The Twitter Archiver app was used to collect the tweets from Twitter, while the comments were manually collected from Facebook. All of the tweets and comments were annotated manually. Initially, a filtering for the collected tweets and comments was considered such as removing the duplication and irrelevant data, before performing the annotation.

3.1.1. Extract the Reviews

In this study, Twitter Archiver app was used to grab the tweets based on some specific vital hashtags, terms, phrases or names of political parties, while a post was in advance published by the author asked people about their opinions on the political situation supported by some important questions. However, the Facebook comments were manually extracted. Table 1 shows the terms on which the reviews were extracted from Twitter and the number of the reviews.

Table 1. Terms and reviews total.

Terms	Number of Reviews
إعادة الأمل (Restoring hope)	135,107
الحرب على اليمن (The war on Yemen)	80,920
الحوثي (Houthi)	90,087
الشرعية في اليمن (Legitimacy in Yemen)	63,730
العدوان السعودي (Saudi aggression	75,483
اليمن (Yemen)	110,650
عاصفة الحزم (Decisive Storm)	120,453
قوات التحالف (Coalition forces)	53,528
Total Terms = 8	**Total Reviews = 729,958**

Twitter Archiver is a Google add-on app that allows the users to easily capture and automatically update all tweets that match certain search terms in a Google Spreadsheet. Initially, installing the app. A new worksheet is directly created. To start fetching the tweets, allow the Google Sheet to access your Twitter account for authorization purposes, and then create a Search Rule. The app grants the user to create rules depending on some specific terms or conditions in which the user wants to get tweets based on. The preliminary obtained tweets are easily saved, and the app automatically pulls in the new tweets when get connected.

On the other side, Figure 2 depicts the nature of the post on Facebook. The author first did dig out in different political websites to know the hot topics of people's debates. Looking at different perspectives gave the author a thoughtful idea to state the contents of the post. Afterward, the author distributed the post on Facebook carried some questions related to the topic. Furthermore, the author asked some of his friends who have many followers to share publicly the post without restrictions to gain more responses. Moreover, the first line of the post illustrates that the post aims for a research purpose, therefore, it was explicitly stated to the participants that by giving their responses they imply their consent in share their views

for research. Although the post was publicly released showing the identification of the participants, some of them preferred to share opinions privately for their safety. However, the total number of responses to the post, whether the once shared by the author, the once shared by selected friends, or the private responses, reached 167 responses. The permission was taken from the concerned party on Twitter to scrape the data using Twitter Archiver App, while there is no such policy on Facebook. Besides, such social media websites allow users to post and grab data for research purposes.

Figure 2. The post copy.

3.1.2. Filter the Reviews

After extracting the reviews based on a specific term, the document saved in Excel file format. Then applied the filtering process such as removing the duplicate tweets/reviews, re-tweets/reviews, and non-Arabic content. Irrelevant reviews, and neutral reviews-which have no sentimental words were excluded. Every aforementioned process was manually performed. The Figures 3 and 4 show the tweets before and after the filtering process.

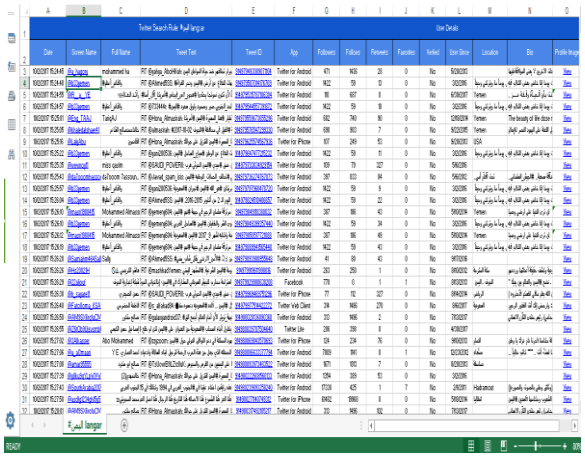


Figure 3. Tweets before filtering.

Figure 4. Tweets after filtering.

3.1.3. Annotate the Reviews

The developed dataset was classified into two classes, positive and negative. The annotation process was manually done by two experts, along with the author. Unlike other domains such as product reviews, people used to clearly express their opinions on whether the product is acceptable, or not. In the case of political reviews, some people use sarcasm opinions, especially when they are not satisfied. Sarcasm is completely the opposite of what the speaker intended to say. For instance, " أيوه عفاش مسك الرئاسة واليمن بالحضيض وأصبح دولة عظمى في عهده". Which means, "Yes, Afash held the presidency while Yemen was in the bottom and then became a superpower in his reign". Another case is, "تضامنهم مع طفلة أبيدت عائلتها بالكامل من قبل طائرات تحالف الوطنية تجري في دمائهم العدوان السعودي شوف". Which means, "Their solidarity with a child whose entire family was annihilated by the planes of the Saudi aggression alliance, see patriotism is running in their blood". Many research studies considered sarcasm reviews as a separate work such as [25]. The dataset of this study contains many sarcasm reviews. The annotators did their best to give a close annotation for such kind of reviews and aimed to be addressed wisely in future work.

3.2. Pre-Processing

Pre-processing is an important step for improving the performance of sentiment analysis by reducing data errors. It is a way to clean up data from unwanted items. Without pre-processing of data, sentiment analysis models may ignore important words and adversely affect the accuracy of the results. The following pre-processing processes are performed in this work.

3.2.1. Process Documents from Data

After created the training dataset, Rapidminer was used to perform the pre-processing step and then build the classification model. Initially, Nominal to Text operator is used to change the type of selected nominal

attribute to text. After conversion, the Tokenize operator is used. The Tokenize operator is responsible to split the text of the tweet into words known as tokens. After token creation, Filter Stopwords (Arabic) operator is used to remove the most common words in the Arabic language. After removal of stopwords, Generate-n-Gram operator is used to identify n of words that appear together several times in the same review. Going back to our developed corpus, many pairs appeared together in the same review such as "عاصفة الحزم" and "عاصفة الأمل", etc., which are the names of military operations carried out by the coalition. Identifying such a couple of words allows the users to parse through the document more smartly. It also addressed the problem of removing the negation words. Because negation words are considered as stopwords, therefore, they were removed, while they are vital words for sentiment analysis, as they can determine the orientation of a review. In this work, 'n' is set to 2, i.e., bi-grams. Where each sequence taken as one token, which in turn led to a slight handling of the problem. The Stem (Arabic) operator is used to reduce different forms of a word to its root or stem form. The "Light Stemming" technique is used. It is designed especially for the Arabic language, and it can remove prefix and suffix of a word.

The same aforementioned steps were applied to the testing dataset with a connection between the two nested operators to match the attributes of the input ExampleSet with the training ExampleSet.

3.2.2. RapidMiner Tool

RapidMiner is a tool developed by the same name company that it can be used to perform the practical tasks of data mining and machine learning. It is used for industrial applications and developer as well as for research purposes. RapidMiner has many operators for machine learning processes, including data pre-processing, classification, and visualization. For data pre-processing, clustering, classification, visualization, and association rules. Also there is a large collection of machine learning algorithms which they are called through the Graphical User Interface. In-text mining classification, process Documents from files creates word vectors from a collection of text stored in multiple files. It also provides multiple-term weighting schemes, and term pruning options [20].

3.2.3. Normalization

Normalization aims to unify a form by mapping all of its variants to a single term-replacing an important part of a term to a common standard that can be easily recognized. Due to the limited capacity of this operator to deal with the Arabic language, this process was manually done via Excel. In this work, some of the normalized terms and letters of the reviews are shown in the Table 2.

Table 2. Normalization rules.

Letters to be normalized	أ	إ	لأ	لإ	عاصفه	جريمه	ضلم
Replace by	ا	ا	لا	لا	عاصفة	جريمة	ظلم

3.3. Feature Selection

Feature selection is the process of selecting a subset of the terms based on its occurrence in the training set. The occurrences of features are calculated to obtain the important features. These features are used as inputs for the classification algorithms. Feature Selection/Extraction is a critical task to be considered in sentiment analysis before building the model. All the extracted features are used as inputs to the classifiers, which lead to get more efficient and improve the performance.

On the other hand, the process documents from data operator has many parameters; the most important one is vector creation, which allows the user to choose a weighing frame. In this study, TF-IDF was selected as a vector creation method.

TF-IDF stands for term frequency-inverse document frequency. It is a text mining technique used to assign weights to terms these have relative importance in a corpus. This weight is a statistical measure used to evaluate importance of a word to a document in the collection or corpus [2].

In this research, two outputs of this method were created, the first one is called Term-Document Matrix which has Row No for rows, the type of the row sentiment, and each word (token or term) found in that document as columns. Each entry in this matrix can be a simple term count, term frequency, or Term Frequency-Inverse Document Frequency (TF-IDF) scores. The second one is called WordList contains Attribute Name which is also the same word (term), the total occurrences in the entire corpus, the number of distinct documents this term occurs in, how many times it occurs as negative and how many times it occurs as positive. Some terms, like names of the political parties or names of the military operations are almost appeared in each review, which they are not sentimental words. Under the concept of the term Frequency (TF), the document which has such words in abundance, in a way that is not wise, will get more weight and leave the more meaningful words with less weight, due to the lack of their occurrences. Therefore, the inverse document frequency normalization reduces the weight of terms, which occurred more frequently in the collection to ensure the matching of documents to be more influenced by that of more discriminative words, which have relatively low frequencies in the collection.

However, Figures 5 and 6 show the Term-document matrix of training dataset and WordList of training dataset, respectively. From the other hand, Figures 7 and 8 show the Term-document matrix of testing dataset and WordList of testing dataset, respectively.

$$tfidf(t) = tf * \log \frac{N}{dfi} \quad (1)$$

Where,

- tfi is term frequency (the occurrences number of a word in a document).
- $df(t)$ is document frequency (the documents number containing the term).
- N is the documents number in the corpus.
- t is the term.
- $tfidf(t)$ is the relative weight of the vector feature.

Row No.	Sentiment	_	_الاحمر	_البيت	_الحوثي	_العدوان	_انطلق	_بارده	_صنعوه	_صلوات
1	Negative	0	0	0	0	0	0	0	0	0
2	Negative	0	0	0	0	0	0	0	0	0
3	Negative	0	0	0	0	0	0	0	0	0
4	Positive	0	0	0	0	0	0	0	0	0
5	Negative	0	0	0	0	0	0	0	0	0
6	Positive	0	0	0	0	0	0	0	0	0
7	Negative	0	0	0	0	0	0	0	0	0
8	Positive	0	0	0	0	0	0	0	0	0
9	Positive	0	0	0	0	0	0	0	0	0
10	Negative	0	0	0	0	0	0	0	0	0
11	Positive	0	0	0	0	0	0	0	0	0
12	Negative	0	0	0	0	0	0	0	0	0
13	Positive	0	0	0	0	0	0	0	0	0
14	Positive	0	0	0	0	0	0	0	0	0
15	Negative	0	0	0	0	0	0	0	0	0
16	Negative	0	0	0	0	0	0	0	0	0

Figure 5. Term-document matrix of training dataset.

Word	Attribute Name	Total Occurences ↓	Document Occurences	Negative	Positive
يمن	يمن	1090	981	709	381
سعد	سعد	694	639	454	240
حرب	حرب	635	557	426	209
حوث	حوث	582	535	435	147
عدا	عدا	578	524	377	201
شرع	شرع	367	326	245	122
عدوان_السعودي	عدوان_السعودي	347	347	237	110
ن	ن	278	252	164	114
حرب_اليمن	حرب_اليمن	277	271	171	106
حلف	حلف	215	205	131	84
له	له	213	190	106	107
عصف	عصف	207	194	79	128
يمني	يمني	204	190	121	83
هزم	هزم	197	192	75	122
شعب	شعب	188	177	120	68

Figure 6. WordList of training dataset.

Row No.	Sentiment	_	_صنعوه	_صلوات	_عودت	_ال	_الـصنعوه	_أول	_أن	_آخر
1	Negative	0	0	0	0	0	0	0	0	0
2	Negative	0	0	0	0	0	0	0	0	0
3	Negative	0	0	0	0	0	0	0	0	0
4	Negative	0	0	0	0	0	0	0	0	0
5	Negative	0	0	0.362	0	0	0	0	0	0
6	Negative	0	0	0	0	0	0	0	0	0
7	Negative	0	0	0	0	0	0	0	0	0
8	Negative	0	0	0	0	0	0	0	0	0
9	Negative	0	0	0	0	0	0	0	0	0
10	Negative	0	0	0	0	0	0	0	0	0
11	Positive	0	0	0	0	0	0	0	0	0
12	Positive	0	0	0	0	0	0	0	0	0
13	Positive	0	0	0	0	0	0	0	0	0
14	Negative	0.180	0	0	0	0	0	0	0	0
15	Positive	0	0	0	0	0	0	0	0	0

Figure 7. Term-document matrix of testing dataset.

Word	Attribute Name	Total Occurences ↓	Document Occurences	Negative	Positive
عصف	عصف	269	244	176	93
هزم	هزم	252	242	166	86
عاصفة_الحزم	عاصفة_الحزم	210	207	137	73
يمن	يمن	99	85	72	27
سعد	سعد	63	60	47	16
ن	ن	42	40	26	16
أمل	أمل	40	33	29	11
عدا	عدا	38	38	31	7
له	له	36	33	11	25
عود	عود	34	33	21	13
حوث	حوث	29	27	15	14
عاصفه_الحزم	عاصفه_الحزم	29	29	20	9
عرب	عرب	28	22	14	14
قتل	قتل	25	24	23	2

Figure 8. WordList of testing dataset.

3.4. Classifications

Classification is a technique used to analyze predefined instances to give a prediction for new instances. In this work, it assigns a predefined class label (positive or negative) for new reviews. In this research, four commonly known classification algorithms were used to build the classifiers to classify the features of textual reviews and predict the sentiment of testing reviews as Positive or Negative namely, Support Vector Machine, NB, KNN, and Decision Tree. Then, the dataset was divided into training and testing parts to train the classifiers. Finally, the classifiers were tested and evaluated based on some performance parameters.

3.4.1. Cross Validation

For the classification purpose of this research, Cross-Validation operator is used, which is a nested operator and has two parts: a training part and a testing part. The training part is used to train a model. The testing part is used to apply the trained model. In the training part, first, Split Data operator is used to split the dataset into training and testing sets. In this work, the dataset was split into 70% of data for the training set and the remaining 30% of data for the testing set. Afterward, the four classification algorithms were utilized one by one.

In the testing part, Apply model operator is used to apply the training set to get a prediction on the testing set. Thereafter, Performance operator is used to show the performance parameters such as Accuracy, Precision, and Recall. Evaluating the performance of all the models was done by settle down 5-fold cross-validation of the Cross Validation operator after trying all numbers from 1 to 10, whereas the F-measure parameter was manually calculated.

Eventually, another Apply model and Performance operators were also applied to the testing dataset. After that, the output of the Cross-Validation is connected to the Apply model of the testing dataset to predict the orientation of the testing dataset reviews.

3.5. Analyzing Results

In this step, the results were analyzed to find out how accurately the performance of the classification models

can determine the polarity of the training and testing datasets. All the four classification models were run on the same training and testing datasets. Then, the results were compared to figure out which classifier performed better according to the performance parameters.

3.5.1. Performance Parameters

The performance parameters are responsible for calculating the accuracy of the classifiers to figure out how accurate the model is and how they can classify the training data. In this research, the results were evaluated by Accuracy, Precision, Recall, and F-measure as they are the measures of the performance.

Accuracy is probably the most intuitive measure of the performance. It is simply the proportion of the correctly expected observations. Precision appears in the proportion of correct positive observations. Recall is also known as sensitivity or true positive rate. It is the proportion of the correctly expected positive events. F-measure is the weighted average of Precision and Recall. Hence, false positives and false negatives are both taken into consideration by this measure.

$$Accuracy = \frac{(TruePos + TrueNeg)}{(TruePos + FalsePos + FalseNeg + TrueNeg)} \quad (2)$$

$$Precision = \frac{(TruePos)}{(TruePos + FalsePos)} \quad (3)$$

$$Recall = \frac{(TruePos)}{(TruePos + FalseNeg)} \quad (4)$$

$$\text{F-measure} = 2 * \frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (5)$$

4. Results and Comparison

All the conducted experiments in this research and a comparison of the outcomes were evaluated based on the performance. Four classification algorithms are used and obtained different results. The algorithm that achieved the higher value of accuracy, recall, and precision, considered as the most efficient-when applying to the training or the testing dataset. All the experiments in this research were carried out using RapidMiner, which described in section 3. The training dataset that comprises of 1307 negative reviews and 693 positive reviews was taken to build the classification models using SVM, NB, KNN, and DT algorithms. Furthermore, the testing dataset, which comprises of 203 negative reviews and 97 positive reviews, was supplied to the models for the prediction purposes. The selected classifiers were applied to the training dataset to build the models. After building the classification models, the testing dataset was supplied to the models for performing the testing task. The obtained results shown in Table 3.

In Table 3, Correctly Classified Instances are those reviews that were negative or positive, and the machine classified them correctly, as they were. On the other hand, Incorrectly Classified Instances are those reviews, which were negative or positive, but the machine classified them incorrectly; shuffled them.

After applying four algorithms to the same training dataset and testing dataset, the following results were obtained. The Table 4 gives a summary comparison of the obtained results based on the performance.

Table 4 shows that, firstly, NB is the fastest classifier and KNN is the slowest classifier. Secondly, Support Vector Machine algorithm showed its superiority over the others in all performance measures for classification of Arabic language.

However, Figures 9, 10, 11, and 12 show the Confusion matrix of SVM, naive Bayesian, KNN and decision tree, respectively.

Table 3. Results of experiments.

	Dataset	Total Number Instances	Correctly Classified Instances	Incorrectly Classified Instances	Percentage
SVM	Training	2000	1813	187	90.65%
	Testing	300	270	30	90.00%
NB	Training	2000	1700	300	85.00%
	Testing	300	248	52	82.67%
K-NN	Training	2000	1740	260	87.00%
	Testing	300	255	45	85.00%
DT	Training	2000	1797	203	89.85%
	Testing	300	262	38	87.33%

Table 4. Comparison of results.

	Time Taken in min	Dataset	Accuracy	Precision	Recall	F-Measure
SVM	2:59	Training	90.65%	91.25	87.96	89.27
		Testing	90.00%	92.00	85.34	87.68
NB	2:22	Training	85.00%	8373	82.76	83.20
		Testing	82.67%	80.53	79.11	79.73
K-NN	7:18	Training	87.00%	86.07	84.86	85.40
		Testing	85.00%	82.90	82.72	82.81
DT	6:52	Training	89.85%	89.93	87.39	88.44
		Testing	87.33%	86.54	83.91	85.01

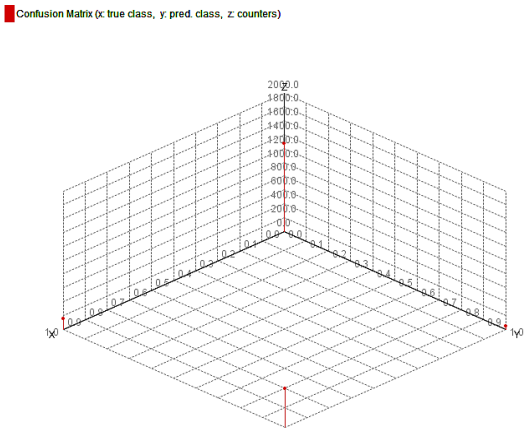


Figure 9. Confusion matrix of SVM on training dataset.

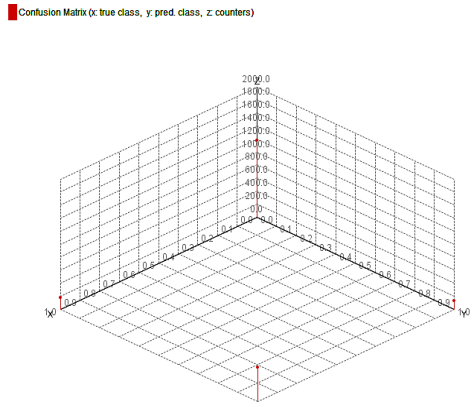


Figure 10. Confusion matrix of naive bayesian on training dataset.

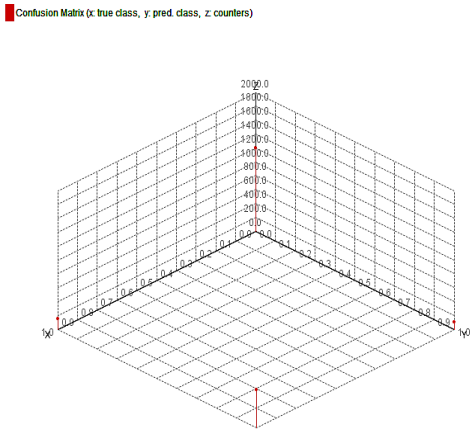


Figure 11. Confusion matrix of K-NN on training dataset.

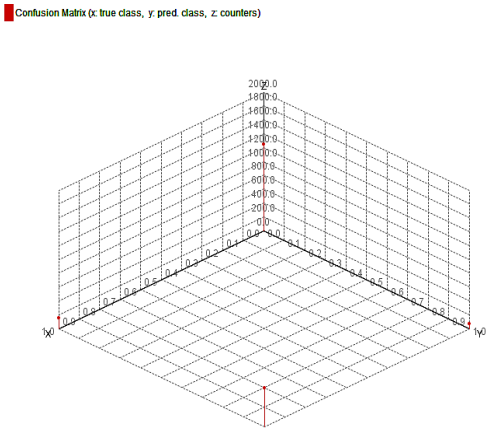


Figure 12. Confusion matrix of decision tree on training dataset.

5. Discussions

NB is the fastest classifier because it assumes that there are no dependencies among attributes. This assumption does simplify the involved computations, which called class conditional independence. Furthermore, it has several free parameters that greatly simplifies the process, such as prior and conditional probabilities. Since it returns probabilities, it becomes easier to apply these results to a wide range of tasks. Moreover, it does not need a large amount of data to start "learning" [10]. However, KNN is the slowest algorithm because of the existed interrelated relationship between the classification time and the data size. As long as the data size is larger, there is a need to calculate large distance, which makes it extremely slow.

SVM algorithm showed its superiority over the others in all performance measures for classification of Arabic language.

Support Vector Machine algorithm outperformed the others in term of Precision on both training and testing datasets, because the training of SVM can minimize a combination of training error (empirical risk) and the probability of incorrect classification unidentified data (structural risk), [18], and the labelled amount of data is more negative than positive.

Finally, the results of this study were affected negatively due to the high accuracy of True Negatives (TN), and the system had a good ability in classifying negative data, but a weak ability in classifying the positive data. Besides, the labelled amount of data is more negative than positive, which in turn led to a low value of True Positive (TP) and F-measure. The same results occurred in a study carried out by [26], which had a high accuracy, but low TP and F-measure.

6. Conclusions

In this modern era, knowing the opinions and attitudes of people is a fast way that enables the concerned people to make an appropriate decision. Hence, instead of wasting time reading and figuring out the polarity of text, we can use the automated techniques for analyzing reviews of people. Therefore, in this research, we introduce an ML-based classification scheme for classifying Arabic reviews. A thoughtful MSA and Yemeni dialects dataset composed of 2000 unique records to classify Arabic reviews has been comprehensively constructed along with a test dataset composed of 300 records used to test the capacity of our scheme. The developed datasets are available to the public for research purposes. Four machine learning algorithms were applied using RapidMiner software. The analyzed results show that Naïve Bayesian is the fastest classifier because it does not need a large amount of data to start learning. However, KNN is the slowest classifier, because its classification time is directly related to the size of data. Finally, the results showed that Support Vector Machine algorithm outperformed the others in all performance measures.

In the future, we will expand the training dataset to give better accuracy and include Neural Network algorithms. The highest consideration for future work is to deal with the sarcasm reviews professionally, as noted in this research has a negative effect on the results. The other thing that should be taken into account is to develop a software that can extract the available online information (reviews) and has a strong

capability to deal with the Arabic language especially for the pre-processing task and all other tasks.

References

[1] Abdulla N., Ahmed N., Shehab M., and Al-Ayyoub M., "Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based," *in Proceeding of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies*, Amman, pp. 1-6, 2013.

[2] Aggarwal C., and Zhai C., *Mining Text Data*, Springer Science and Business Media, 2012.

[3] Alayba A., Palade V., England M., and Iqbal R., "Arabic Language Sentiment Analysis on Health Services," *in Proceeding of IEEE 1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 114-118, 2017.

[4] Al-Ayyoub M., Rihani M., Dalgamoni N., and Abdulla N., "Spoken Arabic Dialects Identification: The Case of Egyptian and Jordanian Dialects," *in Proceeding of 5th International Conference on Information and Communication Systems*, Irbid, pp. 1-6, 2014.

[5] Al-Azani S. and El-Alfy E., "Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text," *in Proceeding of 8th international Conference on Ambient Systems, Networks and Technologies, ANT*, Dhahran, pp. 359-366, 2017.

[6] Al-Harbi O., "Classifying Sentiment of Dialectal Arabic Reviews: A Semi-Supervised Approach," *The International Arab Journal of Information Technology*, vol. 16, no. 6, pp. 995-1002, 2019.

[7] Al-Harbi O., "Using Objective Words in the Reviews to Improve the Colloquial Arabic Sentiment Analysis," *International Journal on Natural Language Computing*, vol. 6, no. 3, pp. 01-14, 2017.

[8] Alhumoud S., Albuhairi T., and Altuwaijri M., "Arabic Sentiment Analysis Using WEKA A Hybrid Learning Approach," *in Proceeding of 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Lisbon, pp. 402-408, 2015.

[9] AL-Sukkar G., Aljarah I., and Alsawalqah H., "Enhancing the Arabic Sentiment Analysis Using Different Preprocessing Operators," *in Proceedings of the New Trends in Information* Technology, Amman, pp. 113, 2017.

[10] Ashari A., Paryudi I., and Tjoa A., "Performance Comparison Between Naïve Bayes, Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 11, 2013.

[11] Bilal M., Israr H., Shahid M., and Khan A., "Sentiment Classification of Roman-Urdu Opinions Using Naïve Bayesian, Decision Tree and KNN Classification Techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 28, no. 3, pp. 330-344, 2016.

[12] Duwairi R., Marji R., Sha'ban N., and Rushaidat S., "Sentiment Analysis in Arabic Tweets," *in Proceeding of 5th International Conference on Information and Communication Systems*, Irbid, pp. 1-6, 2014.

[13] Duwairi R. and Qarqaz I., "Arabic Sentiment Analysis Using Supervised Classification," *in Proceeding of International Conference on Future Internet of Things and Cloud*, Barcelona, pp. 579-583, 2014.

[14] El-Masri M., Altrabsheh N., Mansour H., and Ramsay A., "A Web-Based Tool for Arabic Sentiment Analysis," *Procedia Computer Science*, vol. 117, pp. 38-45, 2017.

[15] Elnagar A., "Investigation on Sentiment Analysis for Arabic Reviews," *in Proceeding of IEEE/ACS 13th International Conference of Computer Systems and Applications*, Agadir, pp. 1-7, 2016.

[16] Han J., Kamber M., and Pei J., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2011.

[17] Itani M., Roast C., and Al-Khayatt S., "Corpora for Sentiment Analysis of Arabic Text In Social Media," *in Proceeding of IEEE 8th International Conference on Information and Communication* Systems, Irbid, pp. 64-69, 2017.

[18] Khan F., Arnold M., and Pottenger W., "Finite Precision Analysis of Support Vector Machine Classification in Logarithmic Number Systems," *in Proceeding of IEEE Euromicro Symposium on Digital System Design*, Rennes, pp. 254-261, 2004.

[19] Liu B., *Sentiment Analysis and Opinion Mining*, Springer Link, 2012.

[20] Liu B., *Web Data Mining: Exploring Hyperlinks*, Contents, *and Usage Data*, Springer Science and Business Media, 2007.

[21] Mahyoub F., Siddiqui M., and Dahab M., "Building an Arabic Sentiment Lexicon Using Semi-supervised Learning," *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 417-424, 2014.

[22] Mostafa A., "An Automatic Lexicon with Exceptional-Negation Algorithm for Arabic Sentiments Using Supervised Classification," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 15, pp. 3662-3671, 2017.

[23] Mukhtar N. and Khan M., "Urdu Sentiment Analysis Using Supervised Machine Learning Approach," *International Journal of Pattern*

Recognition and Artificial Intelligence, vol. 32, no. 02, pp. 1851001, 2018.

[24] Mustafa H., Mohamed A., and Elzanfaly D., "An Enhanced Approach for Arabic Sentiment Analysis," *International Journal of Artificial Intelligence and* Applications, vol. 8, no. 5, pp. 01-14, 2017.

[25] Parveen S., Surnar A., and Sonawane S., "Mining in Twitter: How to make use of Sarcasm to Enhance Sentiment Analysis: A Review," *International Journal of Advanced Research in Computer Engineering and* Technology, vol. 6, no. 6, 2017.

[26] Prabowo D., Setiawan N., and Nugroho H., "A Study of Data Randomization on A Computer Based Feature Selection for Diagnosing Coronary Artery Disease," *Advances in Intelligent Systems*, vol. 53, pp. 237-248, 2014.

[27] Sati B., Ali M., and Abdou S., "Arabic Text Question Answering from an Answer Retrieval Point of View: A Survey," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, pp. 478-484, 2016.

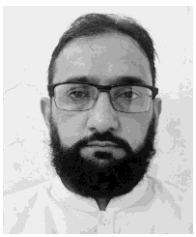
[28] Sohail S., Siddiqui J., and Ali R., "Book Recommendation System Using Opinion Mining Technique," *in Proceeding of International Conference on Advances in Computing, Communications and Informatics*, Mysore, pp. 1609-1614, 2013.

[29] Zubair M., "Survey of Data Mining Techniques for Social Network Analysis," *International Journal of Research in Computer Engineering and Electronics*, vol. 6, no. 3, pp. 01-08, 2014.

Emran Al-Buraihy received the B.Sc. (Hons) degree in information technology from University of Science and Technology Taiz, Yemen, in 2014, and the M.S. degree in information technology from Institute of Business and Management Science (IBMS), The University of Agriculture Peshawar, Peshawar, Pakistan in 2018. He is currently pursuing the Ph.D. degree in Computer Science and Technology at Beijing University of Technology, Beijing, China.

Wang Dan received the B.S. degree in computer application, the M.S. degree in computer software and theory, and the Ph.D. degree in computer software and theory from Northeastern University, China, in 1991, 1996, and 2002, respectively. She is currently a Professor with the College of Computer Science, Beijing University of Technology. Her major areas of interests include trusted software, web security, and big data.

Rafi Ullah Khan received the B.S. degree in computer science from Islamia College Peshawar, Peshawar, Pakistan, in 2007, the M.S. degree in internetworking and digital communication from the Institute of Management Sciences (IMS), Peshawar, in 2010, and the Ph.D. degree in computer science from the Capital University of Science & Technology, Islamabad, Pakistan, in 2020. He has been working as a Senior Lecturer with the Institute of Computer Sciences and Information Technology, The University of Agriculture, Peshawar, Pakistan, since 2011. His research interests include data mining, machine learning, web user privacy, sentiment analysis, and computer networks.



Mohib Ullah received the M.S. degree from Birmingham City University, U.K., and the Ph.D. degree from the Capital University of Science and Technology, Islamabad, Pakistan. He is currently working as a Senior Lecturer with the Institute of Computer Sciences and Information Technology (ICS/IT), The University of Agriculture, Peshawar, Pakistan. He has published 15 research articles in well-reputed journals and international conferences. His research interests include the security and privacy issues associated with computer networks, WSN, and the IoT.