ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse36942020.pdf
https://doi.org/10.30534/ijatcse/2020/36942020

Mood Detection Based on Arabic Text Documents using Machine Learning Methods

Abdelbaset Hussein[1], Mohamed Al Kafri[2], Abdullah A. Abonamah[3], Muhammad Usman Tariq[4]
[1234]Abu Dhabi School of Management, Abu Dhabi, UAE,
ab.alqadi@gmail.com[1], mohad_cs@yahoo.com[2], a.abonamah@adsm.ac.ae[3], m.tariq@adsm.ac.ae[4]

ABSTRACT

Document text classification is utilized for information feature extraction and retrieval as the primary source of digitizing the written information using text classification techniques. Text classification can provide much more information by analyzing the text using machine learning methods. One of the practical applications of text classification is mood detection using machine learning algorithms. Machine learning algorithms allow a practical and beneficial platform for analyzing and detect mood from the text documents. However, there are few applications to analyze the text in Arabic with high accuracy and especially detecting mood using Arabic text documents, messages, or blogs. The main objective of using machine learning algorithms is to detect the accurate mood target value to the given messages. This study focuses on four mood classes (Happy, Sad, Angry, fear). The text analyzed for this study were gathered from some social media and internet blogs. Three kinds of techniques have been identified based on machine learning approaches: Naïve Bays algorithm, k-Nearest Neighbors (KNN) algorithm, and Support Vector Machine (SVM) algorithm. The text was further analyzed for feature selection related to text mining, feature correlation analysis, and information gain. Lastly, splitting the text into training and testing sets for possible models using robust classifiers. After running the selected classifiers for our study, the results showed that Naïve bays classier had the highest achievement in terms of accuracy. Naïve Bays classifiers received 70%, support vector machine classifiers obtained 68.33%, and k-Nearest Neighbors (KNN) algorithm yield 51.67%.

Key words: Document text classification, K-nearest neighbor (KNN), Machine Learning, Support Vector Machine, Naïve Bayes

1. INTRODUCTION

The main idea of this research is to apply various machine learning algorithms that are based on text mining to detect the moods of people while posting texts on social media. The goal is to identify if they are angry, happy, sad, or fearful. In this way, targeted intervention may be designed, including ads, articles, and video clips. The necessity for building such this type of system is due to provide such optimal support for government sectors with regards to detect people's mood. Currently, many social messages, blogs, and internet information are uploaded to the internet without any processing for it [1,2]. It is imperative to construct such a system that can detect people's moods and check what kind of categories in their messages [2,17]. In these circumstances, we used machine learning approaches to analyze people's messages in association with mood detection based on four categories angry, happy, sad, or fearful. We reviewed academic-oriented literature including relevant information from various sources, namely internet resources, articles, and magazines for successful implementation of this system [16, 21]. All information obtained served as a guide and reference for the implementation of this research. Firstly, we draw data from the main components and topics that are connected to the research in the form of hardware and software [6, 10]. The aim was to examine and investigate previous systems that have been developed around study related to mood detection classification problems [1, 12]. It was intention of the research to discuss the components that will be used in developing the proposed system. There are many solutions for detecting moods using face, voice, and text recognition. However, the study detect and track feelings of user in several methods such as selecting mood by offering some available options (e.g., DayRio, worry watch), which assist tracking moods, finding most things that make users so happy and collect a year summary.

Various type of research had been done that are related to this solution, such as [11, 15, 20, 25] have proposed sentiment classification using multiple SVM classifiers. They have assigned the documents into some predefined category and then performed sentiment classification by the polarity of the subject [13, 27]. This task has been achieved with the classifier combination approaches. Unigrams and some Parts of Speech (POS) features have been used to train the review data for generating different classifiers [4]. The classifier selection can be made with the classifier selection method, and they are combined with the help of some combining rules.

The result has been compared with the individual classifier and concluded that combinational classifiers gave a better performance than individual classifiers [3, 4, 41].

1.1 Research Question

We address two basic research questions:

Research Question One: How can mood detection quality of words and sentences be evaluated using machine learning algorithms work best for text mining?

Research Question Two: What forms of classification performance evaluation metrics are useful for mood detection in association with an accuracy percentage?

1.2 Research Aim and Objectives

This research aims to detect the moods of individuals without asking them explicitly, which helps to provide a better study of customers, employees, or citizens in a matter of the level of satisfaction that can be used by private or governmental organizations. This solution can be applied to mobile applications or computer systems with some modifications to the input to utilize most of the sources, including social media, emails, and many other sources of messaging rather than SMS messages.

- In this research, machine learning is used to estimate user's moods by solving two problems: classifying and evaluating words and sentences that been given from text mining (training data).
- We are attempting to detect moods according to sentences rather than the original text mining (testing data).
- This research aims to find the optimal solution based on four categories for specific messages or articles.
- Applied four primary forms of classification performance evaluation metrics to evaluate mood detection in association with an accuracy percentage

Section 2 will provide theoretical background, section 3 methodology, section 4 results, section 5 presents discussion and lastly section 5 conclusion.

2. THEORETICAL BACKGROUND

The research focuses on mood detection of individual through text messages. The solution can be enhanced further in various field such as robotics and automated systems that provide human-like interaction with users to improve the overall working environment and comfortability of users. There are several studies focused on machine learning based on text mining. [7,8,18,19,22] proposed a novel hybrid system of short message service (SMS) classification to predict harm or spam messages using the Naïve Bayes algorithm and Apriori algorithm. The performance of these techniques frequently relies on statistical calculation and analysis of text mining. Authors have treated Naïve Bayes as the most effectual and significant in text mining. It is indicated an essential improvement of inefficient performance and accuracy from the traditional Naïve Bayes classifier utilizing on specific Data Repository called UCL.

[14, 43] applied a large-scale mood analysis detection for document classification based on social media text. The authors attempted to structure their paper into three main stages; firstly, addressing the main issue related to feature selection, feature extraction, and classification method of mood in internet resources. Secondly, they extracted the global feeling from eighteen million documents as large-scale text mining. Finally, the authors derived document classification in association with mood trajectory for an egocentric user used to predict subtle emotion signals in a user-centric manner. To detect the mood classification method, they suggested two main feature sets in psychology that are utilized, presenting that these features sets are practical and efficient, not required training sets. They obtained classification results compared to the previous one (Nguyen, Phung, Adams, & Venkatesh, 2014). The experimental outcomes illustrate that the automatic classifier system can undoubtedly reach the accuracy and effectiveness of the document classification technique.

[5, 9] proposed a new model for a multi-class sentiment classification technique based on the SVM algorithm. To recognize the sentiment in document classification, a confusion matrix of multiple SVM algorithms is structured depending on the outcomes of several SVM algorithms. The findings of the empirical study demonstrated that the suggested technique is to obtain better results than that of the current technique for multi-class sentiment document classification. [23,45] proposed a new method for document classification based on KNN and Naïve Bayes. The collected text mining was based on an XML format. To deal with this situation, they obtained the corpus utilized in their study from TREC Legal Track with more than 3000 text documents and over 20 kinds of classifications approach. Out of the 20 types of classifications, only six are selected with the optimal and robust number of text documents that can feed the algorithm as training text mining. The text document classification text mining is processed using special software called Rapid Miner, and the result illustrates that the optimum value number for k in k-NN happens at k=13. The accuracy and performance on an average scale arrived at % 55.17, which is much better than using Naïve Bayes with % 39.01.

[23,28] presented novel research for emotion detection based on text mining using machine learning approaches. The researchers attempted to overcome problems by implemented performance evaluation metrics using related machine learning classifiers on a benchmark emotion, inserting texting text mining. The empirical results indicated that the performance calculation metrics of various classifiers in regard to recall, precision, and f-score measure. Eventually, an approach with the optimal outcomes is suggested for the emotion classification technique.

[24,30, 42] proposed a new model to detect the characteristics of emotional expression in association with short message text that related to customer comments using machine learning classifiers. The authors inserted an agent notes that connected to the headquarter of the company to provide valuable information about each complement's customer comment, including the delivery of professional knowledge that improves customer fulfilment. To find such optimal solutions, the researchers recommended using a machine learning algorithm based matching analysis approach, namely CAMP, which can assist the company in handling customer comments effectively and efficiently. It is indicated that the CAMP approach can offer such practical supervision for the following-up facility, and the automation can assist speed-up service answer, which increases customer pleasure and keeps customer reliability.

2.1 Data Mining

The data mining classification method is a procedure of classifying various numbers of documents into a predefined category. In this case, it attempts to play an essential and crucial useful role in managing text documents within systematizing the document classification process. There are two signs in machine learning algorithms: supervised learning and unsupervised learning. The former is focused on where the algorithm is being trained with known labelled documents (output target) [29, 31, 48]. The other one where the target values are not available for the document for the classification task. In this scenario, clusters techniques are implemented to observe the natural grouping for documents (messages and articles). In addition to that, each document can be labelled as multi-class documents or single-class documents in supervised learning models. Single class marked document is relatively simple compared to the other one [33, 38, 46].

In this research, we performed classification procedures on different types of materials, available in notepad format, and collected from the internet. We implemented three different machine learning methods based on Naïve Bayes classifier, Support Vector Machine, and K Nearest Neighbors (KNN), to conduct the initial classification techniques for mood detection on Mood text mining [47]. The main idea of applying this kind of research is to assess the performance and accuracy of the first classification method with Precision, Recall, and F-Score.

2.2 Machine Learning

The machine learning technique is considered an effective and efficient tool. It is also can be regarded as a narrow form of artificial intelligence, bestowing on computers the agency to find a suitable solution for data problems without being explicitly programmed [4,5,10]. Such algorithms may be applied to issues related to pattern recognition and classification settings using computational procedures [32,34]. Figure 1 illustrates a general overview of the machine learning classification process.

2.2.1 Classification: The importance of classification procedure may be illustrated as learning a function to have inputs and target value. All data in the text mining is in the form of an object ($\boldsymbol{x}$), while $\boldsymbol{y}$ is a class label assigned to $\boldsymbol{x}$. Also, Classification techniques are processes that are implemented to label data, also known as a descriptive classifier for the new sample that belongs to the predictive model [35].

$$g(x) = w^T x + b \qquad (1)$$

In equation (1), *w* is referred to weight values, *b* is a bias, and T is matrices transpose. The central problem-related c_1 and c_2, the feature vector *x* is assigned to class c_1 if $g(x)>=0$ and to class c_2.

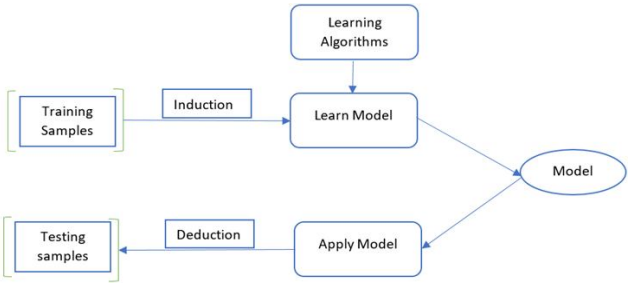

Figure 1: General Machine Learning Classification Approach

2.2 Document Classification based on Naïve Bayes classifier

Naïve Bayes is a classification model that concentrates on Bayes rule, that attempts all the features X1,…, Xn is mutually and provisionally independent given Y. In this case, the value dramatically decreases and simplifies the complexity and the sign of P(X|Y) [19], the main issue with that is to estimate it from the whole training text mining. Let us assume the case where X = (X1, X2), as shown in equation (2) [21].

$$P(X|Y) = P(X1, X2|Y) = P(X1|X2, Y)\, P(X2|Y) = P(X1|Y)\, P(X2|Y) \qquad (2)$$

This equation can be illustrated as shown in the below equation (3).

$$P(X1\ldots..Xn|Y) = \prod P(Xi|Y) \qquad (3)$$

This kind of classifier has been well-known and worldwide widely used due to its simplicity in both the training sets and the target values [36]. Naïve Bayes algorithm permits each feature to participate within the ultimate decision similarly and individually from other features, which is considered

more computational effective and efficient in comparison with document classifiers.

2.3 Document Classification based on Support Vector Machine

Support vector machines (SVM) belongs to a machine learning algorithm and such an essential kind in artificial intelligence that can analyze mood text mining, utilized for classification task and regression [27]. This kind of technique is a class of algorithms that reduces misclassification by a training set, that called maximum margin point [17]. Support Vector Machine was created by two scientists named [37]. Given training data that contain feature and target value; the function refers to document characteristics $(x_1, \ldots, x_n)$ and the target output $\{(y_1, \ldots, y_N), (x_N, y_N)\}$ where $x_i \in input\ features$ and $y_i \in \{class-1, class+1\}$. This algorithm can easily find the optimal solution for the following optimization problem in the next equation (4). Figure 2 shows the optimal separating hyperplane for maximum margin.

$$f(x) = w^T x_i + b \quad (4)$$

$$f(x) = \sum_i \lambda_i y_i (x_i{}^T x + b$$

$$f(x) \geq 1, \quad \forall x \in class\ 1$$

$$f(x) \leq -1, \quad \forall x \in class\ 2$$

$$H = \frac{|g(x)|}{\|w\|} = \frac{1}{\|w\|}$$

W^t refers to the vector weight, while $f(x)$ belongs to features, λ_i refers to a double function, x is the input, y is the target, b bias in association with omega 0. w^t Maximizes the reparability. As demonstrated in the above equation, when vector higher than one, it will belong to a blue circle. Moreover, when values smaller than -1, it will be referred to class number Two (orange circle).

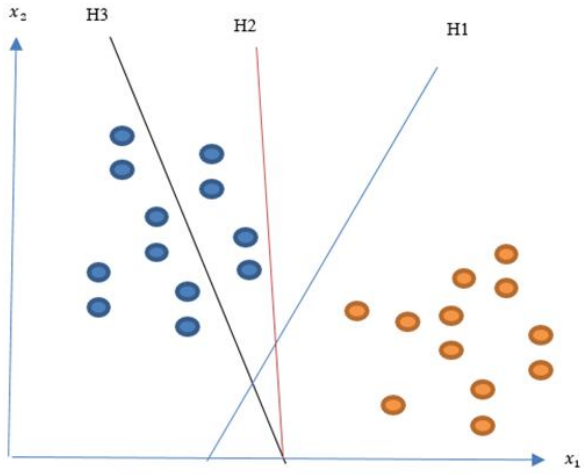

Figure 2: SVM Linearly Separable Set

2.4 Document Classification based on K-nearest Neighbor Classifier

K-Nearest Neighbor Algorithm (KNN) is another kind of artificial intelligence based on machine learning algorithms used in document classification, data mining, statistical pattern recognition, and many more [39]. KNN frequently classifies the training and testing text mining based on the closest attributes space. To provide a clear and correct image of KNN analysis, the process of ranking a new document, which is referred to (query point) samples, is illustrated in Figure 3, which presents a green and yellow sign and the query point with a black circle [28]. The main target and scope are to classify the output of the query point based on its nearest neighbors. Precisely, it is essential to clarify whether a query point can be sorted into an appropriate color. This technique in our experimental study can classify a new object correctly [40].

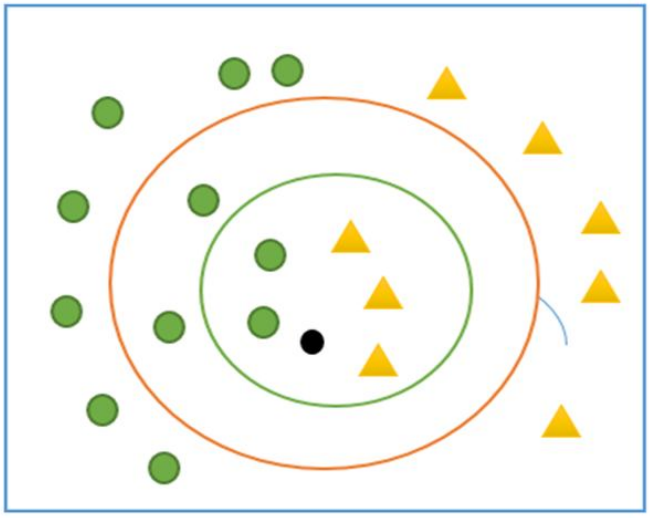

Figure 3: K-nearest neighbor (KNN)

2.4 Word Embedding

One of the most significant issues related to the bag-of-words demonstration is that the selected technique is in charge of discovering which dimensions are considered semantically correlated. To deal with word embeddings representations, the total number of words in each document is characterized as fixed-length embeddings or vectors. There are several models already available to build embeddings. The main target of word embedding is to syntactic and capture semantic symmetries in particular language from unsupervised learning models of documents, such as social media, blogs, internet messages. Vectors present words that happen in the exact context in a nearby manner.

2.5 Term Frequency–Inverse Document Frequency TF-IDF

TF-IDF technique is reflected in how a word is shown in a document (i.e., articles, comments, and blog websites) either in gathering or corpus. Furthermore, this kind of method is

applied as a weighting factor in finding the exact scenario of information retrieval, in particular user modelling and text mining. The TF–IDF value increases with total words available in each text document. It is connected to training and testing sets in the corpus that involve the word that assists in correcting some words appear more frequently.

As mentioned previously, the TF is a calculating of the frequency of the same and exact expression in the same document. Furthermore, since we are dealing with messages, comments, and text documents, everyone is different in length, it can appear many times in long documents than short documents. The equation (5) shows the normalization of the document length to calculate the TF value.

$$TF = \frac{n}{\sum Tn} \quad (5)$$

n is related to items that show in a document, and Tn is the number of terms in the whole document. The second concern related to the IDF where calculation reveals how rare the word exists in a document, as shown in the equation (6).

$$IDF = \log\left(\frac{N}{df}\right) \quad (6)$$

3. METHODOLOGY

The literature review discussed previously highlighted mainly machine learning approaches, document classification text mining using models such as SVM, KNN, and Naïve bays. Feature Selection and Feature Extraction.

3.1 Identification of Data Mining Goal

The primary identification of data mining in our experimental study is divided into three main vital steps. Firstly, the prediction is to determine mood detection based on the attributes in text mining. Secondly, in order to identify patterns in data, it is essential to assess the selected text mining and check how files are related to others and how many mood classes repeated in the same folder. Thirdly, classification was used in our study to Partition text mining into four categories (Happy, Sad, Angry, and Fear). Finally, optimization is also considered to enhance the results using specific kinds of machine learning algorithms.

Machine learning is considered such a useful and optimal method of choice, as they have been proven in several numbers of researchers, but the drawbacks of some of the models are that it is complicated and possesses a nondeterministic polynomial time to solve [17,25,39,40]. To select an appropriate classifier, it is encompassed of trial-and-error processes; therefore, statistical validation is considered suitable to guide that process [22,25,28]. This kind of algorithms can be subjected to underfitting or overfitting, based on the complexity of the method. Moreover, the classification algorithm can be unstable, depending on weights, the timing of training termination. The selection of machine learning approaches has been presented with their statistical and mathematical procedures.

3.2 Proposed Framework

To conduct our experiments using the text mining classification text mining, Figure 4 illustrates the proposed framework of our study. These phases include raw data (text mining), pre-processing, structured data, split text mining, select the suitable model, validation, and presents the outcomes with the best models that generate low error rates.

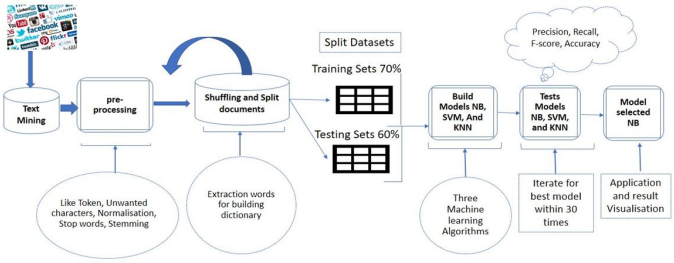

Figure 4: Proposed Framework

Our models NB, SVM, and KNN are considered so powerful because achievement that obtains in association with the training and testing procedure. During our empirical study, the selected statistical methods which affected by outliers offering high and better performance in comparison with Random Forest and Decision tree with a small number of departures that are organized by parametric movements. Generally, according to some researchers mentioned in the literature review section that NB, SVM, and KNN produced a great outcome. Also, with that, usually, this kind of algorithm works better with text mining in comparison with other classifiers like Random Forest and decision tree. These kinds of algorithms are powerful models for the analysis text mining, especially with mood detection, as mentioned previously, that provides such a substantial detection accuracy and performance compare with other models. Therefore, Feature extraction in our model is considered very active due to the selection of specific words related to Mood Detection. Data analysis with many inputs in the selected text mining requires typically a large memory as well as computation power. In this study, we used the embedded approaches for feature extraction, which can easily integrate the model building and the feature selection step. This kind of classifier employs the out-of-bag techniques that can improve the stability of outcomes during training and testing. The selected text mining that was collected from social media, blogs, and internet messages, with 30% of the input for testing and the training received 70% and the remaining percentage, 10% goes for validation, respectively.

3.3 Translation Data

Text mining Data pre-processing method includes knowledge discovery that assists us converted the text mining from undesirable format into an understandable format. It is

crucial to clean the text mining and transform them into the right form to extract robust and efficient results from applying three significant machine learning algorithms. The data gathered from several internet websites. Decreasing the amount of noise and avoiding any missing values are very crucial to acquire a robust quality of accuracy and performance. Inconsistent, Inaccurate, incomplete, and contaminated data analysis can lead to having poor quality and shallow results. Incorrect text mining means having misleading results; this may occur due to data entry errors [35]. Nevertheless, the main target of preparation and pre-processing technique is to indicate the insufficiencies and limitations of text mining.

3.4 Tokenization

In our research, Tokenizer documents can divide into three critical categories of words, phrases, or whole sentences. All the collected messaged and internet blogs in each record have been used in notepad split it into words, phrases, and complete sentences. To avoid any extra punctuation, this kind of technique is used to discard it.

3.5 Stop-words removal

Stop-words are considered such a significant part in pre-processing, which removes articles conjunctions, and prepositions in any documents. The main reason behind that, these kinds of words do not affect the performance or accuracy of the machine learning algorithms. Removing the Stop-words from any articles is to assist our program in identifying the important in words. In our research, words eliminated (i.e., a, an, and, and the).

3.6 Stemming

The natural shape of the Arabic language is a bit difficult to stem in some cases. The light stemmer is fast and straightforward, and no grammatical analysis to determine the root is required.

3.7 Elimination of Unwanted Characters

The best procedure is to be applied in machine learning before any testing is to eliminate string or unwanted characters Arabic numbers, punctuation, marks, and hashtags. Several regular expressions are utilized, as shown in Table 1. As we are dealing with the Arabic language in our experimental study, it is essential to remove everything unrelated to the selected language. In this case, we removed everything that connected to the English language as considered intentional communication and widespread. Numbers and Punctuations also were removed from our text mining.

Table 1: Regular expressions list

No	*Regular expression*
1	[a-zA-A0-9] +
2	[0- 9] +
3	[# - _ ; ? , ! , . ؟ ,]
4	#(&*)+

3.8 Feature Selection

Feature selection is considered such a vast field in pattern recognition and statistical techniques, in association with machine learning models. The idea behind this approach is to select a proper subset from the text document text mining by removing irrelevant features with less critical information, as mentioned in the pre-processing section. This kind of approach is capable of eliminating separate elements to offer a correct decision that could make accidental associations in learning algorithms. In our research, a feature selection technique has been implemented to decrease the unnecessary features before passing to assess the algorithms. To get high classification accuracy and performance, the feature selection considerably enhanced the outcomes of text mining. One of the main benefits is to avoid over-fitting in the algorithms. Two vital processes require to be taken into consideration when choosing the accurate feature subsets. First, it is needed to seek the best feature subsets based on the goodness of aim function, as shown in Figure 5. After that, select the feature subsets with the real purpose. Once the module is accomplished, the last feature subsets are ready to apply machine learning models.

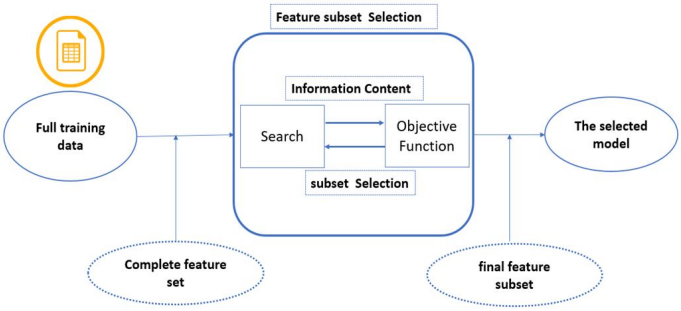

Figure 5: Feature Selection Process

3.9 Machine Learning Model Parameters

Machine learning algorithms with different parameters taken into consideration in our experimental study, which are considered very efficient and effective to obtain good accuracy and performance. Table 2 describes the complete settings for the Naïve Bays model. We selected base-level classifiers to assess the confusion matrices. This model involves Precision, F1 Score, and recall. Naïve Bays, SVM, and KNN will be evaluated individually and discussed in the following section. Furthermore, the performance evaluations are for data drawn from some probability distributions, particularly for deliveries that are not standard. RFC and NN are powerful models for the analysis of SCD text mining, as has been proven for this domain to offer sharp prediction accuracy and performance in comparison with other classifiers. This type of classifier/algorithm employs the out-of-bag method instead of cross-validation, which enhances the stability of results during the training and testing process. A good relationship between input features and target values is discovered during the development process. The text mining was moderate in size, with 20% of the input features randomly selected for testing and the remaining percentages of 70% and 10% used for training and validation, respectively. In this context, the

test set errors are averaged, and the procedure was repeated several times.

Generally, RFC preserves the appealing attributes of decision trees, for instance, handling of redundant/irrelevant descriptors, numerous mechanisms of action, the capability to deal with both regression and classification, and the ability to handle various kinds of descriptors simultaneously. This model was much faster concerning the training procedure, in comparison to the ensemble techniques. A critical reason that RF with NN produced the highest performance is that the model did not have the issue of over-fit, and most importantly, did not require guidance. Additionally, this approach can adequately estimate the significance of features, specifically for classification. Some of the variables are mislabelled for our text mining; the algorithm can handle and detect such missing values, in addition to operating effectively on unbalanced and categorical data, which is less viable for other classifiers, such as SVMs. With the integration of accuracy and efficiency in addition to the useful analytical techniques, the RF and NN algorithms constitute a viable and effective technique for the multi-source classification of SCD text mining, where no suitable statistical algorithms are available. The results gained from the empirical investigation into the use of various types of machine learning models show that the chosen text mining exhibit significant non-linear relationships, presenting a challenge for the test models. Of the combined classifiers understudy, the NN-RFC outperformed the other models, as illustrated, demonstrating capability in fitting during the testing phase.

Table 2: NB Parameters

Parameters	*Value*	*Details*
Alpha	1	Utilizing Laplace for smoothing feature
Fit prior	True	To evaluate and learn prior class probabilities.
Class prior	None	Prior probabilities of the classes.
Binarize	0	The threshold for binarizing of sample features used by Bernoulli
Norm	False	the second normalization of the weights is performed by complement

Table 3 describes the complete parameters for the SVM model.

Table 0: SVM Parameters

Parameters	*Value*	*Details*
C	1.0	Penalty parameter
Coef0	0.0	Independent term in kernel function. It is only significant in the poly kernel and sigmoid kernel
Max iteration	-1	No limit
Tol	0.001	Tolerance for stopping criterion
Cache size	200	A cache of kernel size in MB
Decision function shape	OVO	Decision function which has shape (samples, classes * (classes - 1) / 2).
Kernel	Linear	

Table 4 describes the complete parameters for the KNN model.

Table 4: KNN Parameters

Parameters name	*Value*	*Details*
N_ neighbors	5	Number of neighbors to use
Weight	uniform	All points in each neighborhood are weighted equally.
P	2	Power parameter for the Minkowski metric. When p = 2, this is equivalent to Euclidean distance.
Metric	Minkowski	The default metric is Minkowski, and with p=2, it is equivalent to the standard Euclidean metric.
metric_params	None	Additional keyword arguments for the metric function.
n_jobs	None	The number of parallel jobs to run for neighbor's search.

3.10 Performance Evaluation Metrics

The evaluation was carried out in a confusion matrix that includes such an important feature to evaluate the accuracy and performance of each model. To evaluate each model, we utilized the confusion matrix frequently. Table 5 explains in equations how the performance evaluation measurements have been calculated.

Table 5: Metric calculations

Type	*Calculation*
Recall	$TP/(TP+FN)$
Precision	$TP/(TP+FP)$
F1 Score	$2 * (Precision * Recall)/(Precision + Recall)$

4. RESULTS

In this section, the whole simulation outcomes and analysis of the document classification text mining are represented. Three important single classifiers have been implemented to evaluate the proposed models in association with performance calculation metrics shown in the previous sections. Naïve Bays, Support Vector Machine and K-nearest neighbor algorithm were chosen based on the unsupervised learning models due to the unknown label of the collected text mining. Machine learning algorithms offer different essential properties, like universal approximation, non-linear mapping, and parallel processing. Moreover, experiments are carried on text mining data to assess various classifiers

illustrated in the methodology. The text mining data was gathered from several social media and internet blogs. Table 6 shows the text mining data that was collected. The text mining data contained 199 samples, with one target value describing with four main categories (Happy, Sad, Angry, and fear). An appropriately enormous amount of text mining was managed to gather to apply it with the machine learning algorithm. The main decision creates a trade-off since our training and testing was limited to 199 cases. Figure 6 shows a description of the mood text mining data, and Table 7 illustrates the proportion of training and testing sets.

Figure 6: Text Mining Data

Table 6: Training Data Set Size

Class	*Numbers*
Angry	50
Fear	50
Happy	50
Sad	49

Table 7: Text Mining Training and Testing proportion

Class	*Training*	*Testing*
Angry	37	13
Fear	34	16
Happy	33	17
Sad	35	14

Figure 8 shows a histogram of the four classes that indicates the total distribution is considerably skewed in favor of the text mining classification. This experimental research was conducted with the study objectives and aim and find a suitable solution for mood problems. Hence, this text mining contains some or just a few errors, which require cleaning. In this study, we built on the multi-class issue due to the text mining comprising four classes.

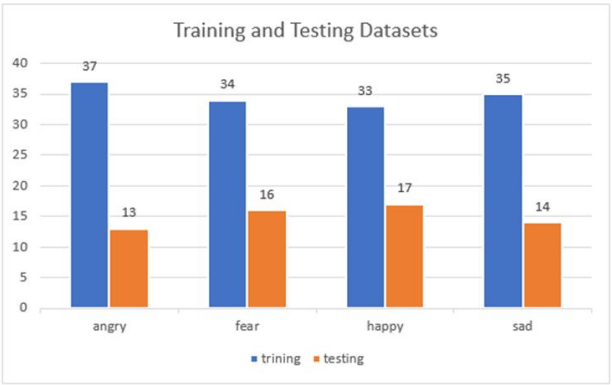

Figure 8: Histogram of Training Classes

4.1 Research Question One

We applied three important classifiers that evaluate the mood detection for Arabic text mining to find a suitable solution for the research question based on machine learning algorithms. The following three subsections answer our first research question.

4.1.1 Naïve Bays classifier Results

The initial performance calculation was performed on text mining, which involves 199 observations. The experiential research was conducted utilizing algorithms in regard to Naïve Bays classifier. Each model was assessed by using the evaluation metrics to calculate the performance classification method. The selected text mining was divided into two main relevant sections: a training set and a testing set, which were randomly chosen with iteration at each run. The results from our study are shown in Table 8. During the model construction, we discovered, class Fear during precision was performed with an excellent outcome in association with the best accuracy with a rate of 0.80. In addition to that, class Angry achieved the best percentage at 0.85. Finally, we discovered after running our simulation using a Python programming language that considers uncommercial and free installation, the F-score with class angry outperformed all the other classes with 0.81.

Table 8: Performance evaluation metrics with Naïve Bays classifier

Class	*Precision*	*Recall*	*F-score*
Angry	0.79	0.85	0.81
Fear	0.80	0.75	0.77
Happy	0.73	0.65	0.69
Sad	0.50	0.57	0.53

4.1.2 Support Vector Machine Results

SVM is considered a binary classifier that takes a set of variables as input and classifies each input into two categories only. This model maps all the selected text mining and convert all n-dimensional sample values into an h-dimensional attribute. Besides, the new attribute is classified through the construction of a specific linear approach. Considering that the main contribution of this research is to recognize geometrical configurations with four classes with mood text mining. In the proposed model, we applied SVM using various kinds of kernels. The results illustrated in Table 9. SVM yields better performance evaluation metrics based on precision with happy class 0.85, where recall offers the best rate with angry class 0.92, finally happy class in association with F-score obtain 0.73.

Table 9: Performance Evaluation Metrics with SVM Classifier

Class	*Precision*	*Recall*	*F-score*
Angry	0.52	0.92	0.67
Fear	0.83	0.62	0.71
Happy	0.85	0.65	0.73
Sad	0.67	0.57	0.62

4.1.3 KNN Classifier Results

The main target of using several classifiers to compare and assess each classifier, which can perform well. Each class in our text mining is organized to the specific classes (angry, happy, sad, and fear). Such a vital division is handled to offer proper class representation in terms of the training and testing text mining. The decision in this model presents a trade-off since the collected text mining was limited to about 199 documents. To deal with this number of examples, the simulation classification outcomes using KNN shown that the proposed classifier yields the lowest outcomes using the performance calculation methods. KNN performed well during the training and testing stage and provided acceptable results, as shown in Table 10. The figures for both training and testing sets are shown in figure 11. Moreover, the performance obtained with class happy during precision 0.67, where angry and fear class acquired the same results with 0.62, and finally F-score with angry class offer highest results compare with other classes with rate 0.59.

This kind of model presents a new method to utilize a different specific number of K to check the calculation performance metrics techniques during the training and testing process. This research used K-nearest neighbors with 50 k-nearest neighbors. In addition to that, during the investigation task, it explored the KNN with 10 K produced such acceptable results, but overall this algorithm produced less accuracy in comparison with other approaches. To discover the closest similar points, it is crucial to find the distance between each point utilizing distance measures such as Euclidean distance. KNN is used these three necessary steps calculate distance, find closest neighbors, and vote for labels.

Table 10: Performance Evaluation Metrics with KNN Classifier

Class	*Precision*	*Recall*	*f-score*
Angry	0.57	0.62	0.59
Fear	0.45	0.62	0.53
Happy	0.67	0.59	0.52
Sad	0.33	0.21	0.26

4.2 Research Question Two Results

The obtained results and the methodology section in our study have already been reflected in the results section. In this piece of research, data science is implemented four classes extracted from 199 documents for mood detection based on classification text mining. The reason behind our selection of Naïve bays method is dominant during the training sets and testing sets process. The accuracy of Naïve bays achieved 70 %, where SVM received 68.33 %, and finally, KNN received the lowest results with a rate of 5.67 %, as shown in Table 11. Our study presented such great statistical methods with robust performance. Overall, the body of outcomes that acquired evaluate the potential of mood detection for the classification text mining. To obtain a satisfactory result, it is so vital to select the appropriate model. The Naïve Bays classifier successfully responded to text mining and the potential use by the psychologist people.

Table 11: Models Comparison

Model	*Precision*	*Recall*	*F-score*	*Accuracy*
NB	0.71	0.70	0.70	70 %
SVM	0.73	0.68	0.69	68.333%
KNN	0.51	0.51	0.51	51.666%

5. DISCUSSION AND RESULTS INTERPRETATION

The result interpretation using machine learning classifiers based on the performance evaluation metrics technique during the training and testing process. This piece of research used four classes (happy, sad, angry, and fear). During our investigation process, we found that the Naïve Bays classifier produced the optimal results in comparison with SVM and KNN. This outcome was expected, as mentioned by many types of research, that Naïve Bays works best with text mining. For validating results, we used 199 documents of our text mining, including four classes. This thesis concentrated on the multi-class label classification issue to get training and testing sets for all the single models along with other performance calculations. We also offer further performance visualizations using specific plots in Figure 9. The line graph in the scheme provides a significant visual comparison across the classifier based on training and testing sets.

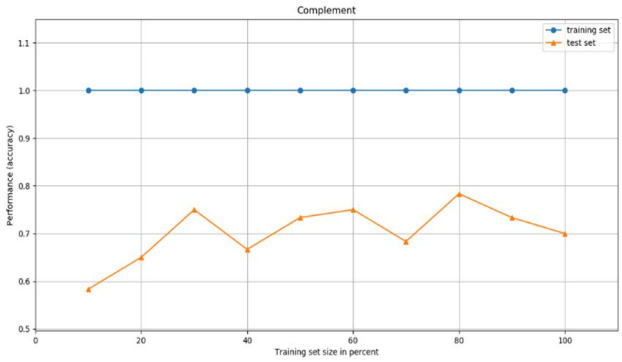

Figure 9: Training and Testing Results using Naïve Bays

In our Arabic text mining, the data points in the SVM classifier are considered not linearly separable due to the four classes with multi-class matters. It is crucial to apply a non-linear mapping (φ) technique within dimension universe to obtain such excellent results in terms of accuracy with a multi-class problem, [12]. This research conducted experimental simulation based on SVM with several iterations at each run using Python programming language. We have implemented SVM using such a robust parameter after that explored deeply the effect of normalization with the current classifiers on the evaluation calculation based on training and testing text mining. We carried out a further empirical study using the training and testing sets, as shown

in figure 10, presenting this model is significantly less able to be classifying our text mining data. It is noticed that the testing line graph is gradually increased to hit high accuracy and performance.

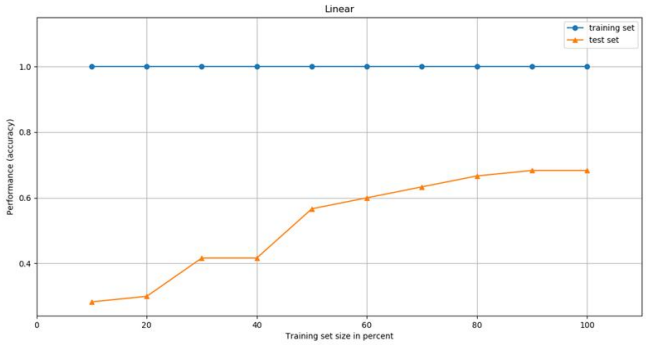

Figure 10: Training and Testing Results Using SVM

The line graph for the KNN model during training sets is ranged from 0.5 to 0.8 for an average of four multi-classes. The classifier outcomes were presented such a poorest in comparison to Naïve bays and SVM due to the KNN frequently performing not optimal more than two classes and performing well with two classes. In Figure 11 of a line graph, it is pointed out that training text mining started at 0.50 then gradually reached just below 0.8. At the same time, testing sets based on experimental results show considerable improvement that reached to 0.65. It is significant to obtain good accuracy and performance within mood detection based on Arabic text mining.

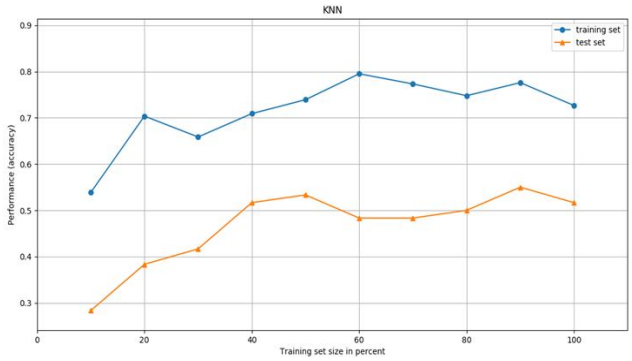

Figure 11: Training and Testing Results using KNN

5.1 Model Reflection

The method reflection of each model selected for our experimental study is calculated through a threshold based on four classes (Happy, Sad, Angry, and fear) to select ultimate class membership of a specific objective. Our models consist of ensamples, which known as training and out-of-sample, as called testing. To evaluate each model individually using the performance evaluation calculation, it is so essential to use accuracies such as recall, precision, F1 score, and overall classification accuracy. Also, it is significant to represent the results in a line graph. Figure 12 shows the total reflection of average and precision per model.

The objective of this study was to use filtering methods for feature selection related to text mining, feature correlation analysis, and information gain, lastly, split the text mining into and training sets and testing sets possible models using robust classifiers. After running the selected classifiers in our experimental study, which considered as a single model, the results implied that Naïve bays classier presented the highest achievement in terms of accuracy. Naïve Bays received 70%, support vector machine classifiers obtained 68.33%, and k-Nearest Neighbors (KNN) algorithm yield 51.67%.

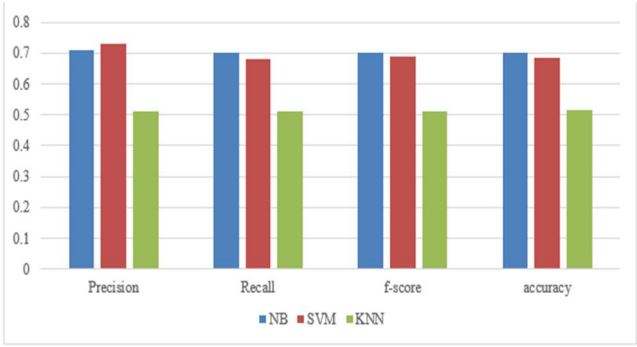

Figure 12: Overall Reflection of Average and Accuracy per Model

6. CONCLUSION AND FURTHER RESEARCH

Various types of machine leanings approach with more details of supervised and unsupervised machine learning have been discussed in this research. Then we presented brief about document classification and some essential statistical tools to document classification in addition to previous studies that used Naïve bays, SVM, and KNN to classify text mining. On the other side, we discussed the methodology and experimental setup of the research. The processing method was explained in detail, including text data collection and pre-processing, data integration and normalization, including tokenization, and stop words. Feature selection also has been discussed for selecting suitable features that can be used for training and testing processes. Three single classifiers have been addressed with a full description of each model, including their parameters. It is so important also to include the proposed framework and the implementation of the experimental setup to find suitable solutions for some of the matters in association with the text mining problems. Currently, no research has been applied regarding text mining based on four categories (Angry. Fear, happy, and sad) in association with various kinds of messages collected from different social media and blogs with their mood as a target value. Text mining has been raised to develop the current models or produce different ones by using different classifiers that generate as few errors as possible. Although text mining used in our experimental research is collected from various kinds of internet resources, it can be utilized to offer accurate information for psychologists for each person individually. Text mining comprises several different features

that help to detect the correct mood of people. Using several machine learning algorithms like Naïve Bays, KNN, and SVM techniques as robust learners, using as baseline classifiers, we trained these algorithms on the collected text mining. We applied testing sets to evaluate them according to their accuracy and performance. In addition to that, it is very vital to get results with very low error arrearage. To improve efficiency and performance, we applied several techniques; feature selection, feature extraction, and confusion matrix, which comprises learning to obtain sophisticated outcomes.

5.1 Further Research

With the success that occurs in our research, we consider some suggestions as to work directions to proposed machine learning algorithms further using single classifiers. Further research is fundamental to confirm our outcomes. In this matter, we suggest some possible extensions to mood detection research, as discussed below:

- Utilize the global optimization classifier, such as a genetic algorithm, to discover more comprehensively related machine learning techniques.
- Some reactions according to the detected mood in many ways like playing music, calling concern friends, or suggesting some treatment actions in critical issues (i.e., emergency call).
- Collect a massive amount of text mining from different internet sources to evaluate and obtain better performance and accuracy.
- Although the proposed classifier is concentrating on some specific parameters of the classification metrics, it is highly recommended to use the area under the ROC curve (AUC) and Receiver operating characteristic.

REFERENCES

1. Abdullah, M., Hadzikadicy, M., & Shaikhz, S. (2018, December). SEDAT: sentiment and emotion detection in Arabic text using CNN-LSTM deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 835-840). IEEE. https://doi.org/10.1109/ICMLA.2018.00134
2. Abd, D. H., Sadiq, A. T., & Abbas, A. R. (2020). *Classifying Political Arabic Articles Using Support Vector Machine with Different Feature Extraction*, Cham.
3. Adankon, M. M., & Cheriet, M. (2009). Model selection for the LS-SVM. Application to handwriting recognition. *Pattern recognition, 42*(12), 3264-3270. doi:https://doi.org/10.1016/j.patcog.2008.10.023
4. Ahmad, M. A. B. (2013). *Mining health data for breast cancer diagnosis using machine learning*: University of Canberra.
5. Akour, M., Al Qasem, O., Alsghaier, H., & Al-Radaideh, K. (2019). The effectiveness of using deep learning algorithms in predicting daily activities. *International Journal*, *8*(5). https://doi.org/10.30534/ijatcse/2019/57852019
6. Al-Khatib, A., & El-Beltagy, S. R. (2017, April). Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 105-114). Springer, Cham.
7. Al-Mahdawi, A., & Teahan, W. J. (2019). *Automatic emotion recognition in English and Arabic text* (Doctoral dissertation, Bangor University).
8. Al-Saqqa, S., Abdel-Nabi, H., & Awajan, A. (2018, July). A survey of textual emotion detection. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)* (pp. 136-142). IEEE.
9. Alsharif, O., Alshamaa, D., & Ghneim, N. (2013). Emotion classification in Arabic poetry using machine learning. *International Journal of Computer Applications*, *65*(16).
10. Al-Omari, A., & Abuata, B. (2014). Arabic light stemmer (ARS). *Journal of Engineering Science and Technology*, *9*(6), 702-717.
11. Alm, C. O., Roth, D., & Sproat, R. (2005). *Emotions from text: machine learning for text-based emotion prediction.* Paper presented at the Proceedings of the conference on human language technology and empirical methods in natural language processing.
12. Asghar, M. Z., Subhan, F., Imran, M., Kundi, F. M., Shamshirband, S., Mosavi, A., . . . Varkonyi-Koczy, A. R. (2019). Performance evaluation of supervised machine learning techniques for efficient detection of emotions from online content. *arXiv preprint arXiv:1908.01587*.
13. Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). paper recommender systems: a literature survey. *International Journal on Digital Libraries, 17*(4), 305-338.
14. Christopher, A. A., & alias Balamurugan, S. A. (2014). Prediction of warning level in aircraft accidents using data mining techniques. *The Aeronautical Journal, 118*(1206), 935-952.
15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.
16. Deconinck, S. (2010). Artificial intelligence a modern approach.
17. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.
18. El Gohary, A. F., Sultan, T. I., Hana, M. A., & El Dosoky, M. M. (2013). A computational approach for analyzing and detecting emotions in Arabic text. *International Journal of Engineering Research and Applications (IJERA), 3*, 100-107.
19. Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), 102121.

20. Hric, M., Chmulík, M., & Jarina, R. (2011). *Model parameters selection for SVM classification using Particle Swarm Optimization.* Paper presented at the Radioelektronika (RADIOELEKTRONIKA), 2011 21st International Conference.
21. Hussain, A. J., Fergus, P., Al-Askar, H., Al-Jumeily, D., & Jager, F. (2015). Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women. *Neurocomputing, 151*, 963-974.
22. Inbarani, H. H., Azar, A. T., & Jothi, G. (2014). Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Computer methods and programs in biomedicine, 113*(1), 175-185. https://doi.org/10.1016/j.cmpb.2013.10.007
23. Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer, 29*(3), 31-44.
24. Khalaf, M., Hussain, A. J., Al-Jumeily, D., Keenan, R., Keight, R., Fergus, P., & Idowu, I. O. (2015). *Applied Difference Techniques of Machine Learning Algorithm and Web-Based Management System for Sickle Cell Disease*. Paper presented at the Developments of E-Systems Engineering (DeSE), 2015 International Conference on.
25. Khalaf, M., Hussain, A. J., Keight, R., Al-Jumeily, D., Fergus, P., Keenan, R., & Tso, P. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. *Neurocomputing*. doi:http://dx.doi.org/10.1016/j.neucom.2016.10.043
26. Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection* (Vol. 1): MIT press.
27. Kaur, J., & Saini, J. R. (2014). Emotion detection and sentiment analysis in text corpus: a differential study with informal and formal writing styles. *International Journal of Computer Application, ISSN*, 0975-8887. https://doi.org/10.5120/17712-8078
28. Li, S., & Zong, C. (2008). *Multi-domain sentiment classification.* Paper presented at the Proceedings of ACL-08: HLT, Short Papers.
29. Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. *Information Sciences, 394-395*, 38-52. doi:https://doi.org/10.1016/j.ins.2017.02.016
30. Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2014). Mood sensing from social media texts and its applications. *Knowledge and Information Systems, 39*(3), 667-702. doi:10.1007/s10115-013-0628-8
31. Ottom, M. A., Alawad, N. A., & Nahar, K. M. (2019). Classification of Mushroom Fungi Using Machine Learning Techniques. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(5), 2378-2385. https://doi.org/10.30534/ijatcse/2019/78852019
32. Pandya, S. S., & Kalani, N. B. (2019). Review on text sequence processing with use of different deep neural network model. *Int. J. of Advanced Trends in Computer Science and Engineering*, *8*(5), 2224-2230. https://doi.org/10.30534/ijatcse/2019/56852019
33. Parvin, H., Alizadeh, H., & Minaei-Bidgoli, B. (2008). *MKNN: Modified k-nearest neighbor.* Paper presented at the Proceedings of the World Congress on Engineering and Computer Science.
34. PhysioNet. (2012). The Term -Preterm EHG Database (TPEHG- DB). *physionet.org*.
35. Putra, R. R., Johan, M. E., & Kaburuan, E. R. (1856). A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia. *International Journal of Advanced Trends in Computer Science and Engineering*, *1860*. https://doi.org/10.30534/ijatcse/2019/07852019
36. Rabie, O., & Sturm, C. (2014). Feel the heat: Emotion detection in Arabic social media content. In *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)* (pp. 37-49).
37. Rushdi-Saleh, M., Martín-Valdivia, M. T., Lopez, L. A. U., & Perea-Ortega, J. M. (2011, September). Bilingual experiments with an arabic-english corpus for opinion mining. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (pp. 740-745).
38. Salloum, S. A., AlHamad, A. Q., Al-Emran, M., & Shaalan, K. (2018). A survey of Arabic text mining. In *Intelligent Natural Language Processing: Trends and Applications* (pp. 417-431). Springer, Cham.
39. Sable, S., & Kalavadekar, P. (2016). SMS Classification Based on Naïve Bayes Classifier and Semi-supervised Learning. *International Journal of Innovations in Engineering Research and Technology*, *3*(7).
40. Sánchez A, V. D. (2003). Advanced support vector machines and kernel methods. *Neurocomputing*, *55*(1-2), 5-20. https://doi.org/10.1016/S0925-2312(03)00373-4
41. Satapathy, S., & Bhagwani, S. (2012). Capturing Emotions in Sentences. *Retrieved March, 15*, 2012.
42. Shahabi, C., Kolahdouzan, M. R., & Sharifzadeh, M. (2003). A Road Network Embedding Technique for K-Nearest Neighbor Search in Moving Object Databases. *GeoInformatica*, *7*(3), 255-273. doi:10.1023/a:1025153016110
43. Syamala, M., & Nalini, N. J. (2019). A Deep Analysis on Aspect based Sentiment Text Classification Approaches. *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, *8*(5), 1795-1801. https://doi.org/10.30534/ijatcse/2019/01852019
44. Tan, P.-N. (2006). *Introduction to data mining*: Pearson Education India.
45. Ting, S., Ip, W., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification.

International Journal of Software Engineering and Its Applications, *5*(3), 37-46.

46. Wang, T.-Y., & Chiang, H.-M. (2011). Solving multi-label text categorization problem using support vector machine approach with membership function. *Neurocomputing*, *74*(17), 3682-3689.
47. Wei, Q., Shi, X., Li, Q., & Chen, G. (2020). *Enhancing Customer Satisfaction Analysis with a Machine Learning Approach: From a Perspective of Matching Customer Comment and Agent Note.* Paper presented at the Proceedings of the 53rd Hawaii International Conference on System Sciences.
48. Wei, Z.-S., Han, K., Yang, J.-Y., Shen, H.-B., & Yu, D.-J. (2016). Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing*, *193*, 201-212. doi:http://dx.doi.org/10.1016/j.neucom.2016.02.022