ISSN 2278-3091
Volume 10, No.2, March - April 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse391022021.pdf
https://doi.org/10.30534/ijatcse/2021/391022021

A Two-Stage Machine Learning Classification Approach to Identify Extremism in Arabic Opinions

Emad M. Al-Shawakfa[1], Hind H. Husni [2]
[1]Yarmouk University, Jordan,, Shawakfa@yu.edu.jo
[2]Yarmouk University, Jordan, hindmadi1993@gmail.com

ABSTRACT

The increased usage of the Internet and social networks allowed and enabled people to express their views, which have generated an increasing attention lately. Sentiment Analysis (SA) techniques are used to determine the polarity of information, either positive or negative, toward a given topic, including opinions. In this research, we have introduced a machine learning approach based on Support Vector Machine (SVM), Naïve Bayes (NB) and Random Forest (RF) classifiers, to find and classify extreme opinions in Arabic reviews. To achieve this, a dataset of 1500 Arabic reviews was collected from Google Play Store. In addition, a two-stage Classification process was applied to classify the reviews. In the first stage, we built a binary classifier to sort out positive from negative reviews. In the second stage, however we applied a binary classification mechanism based on a set of proposed rules that distinguishes extreme positive from positive reviews, and extreme negative from negative reviews. Four major experiments were conducted with a total of 10 different sub experiments to fulfill the two-stage process using different X-validation schemas and Term Frequency-Inverse Document Frequency feature selection method. Obtained results have indicated that SVM was the best during the first stage classification with 30% testing data, and NB was the best with 20% testing data. The results of the second stage classification indicated that SVM has scored better results in identifying extreme positive reviews when dealing with the positive dataset with an overall accuracy of 68.7% and NB showed better accuracy results in identifying extreme negative reviews when dealing with the negative dataset, with an overall accuracy of 72.8%.

Key words : Machine Learning, Arabic Sentiment Analysis, Opinion Mining, Extremism.

1. INTRODUCTION

With the emergence of many products from different vendors, different application software suitable for our daily needs, or even different movies to watch, it became more and more difficult to choose among them. To do that, we usually tend to ask others, or even read a ton of different critiques; a process that might be time consuming. The automatic task for investigating these reviews, opinions, expressions, attitudes, or even behaviors and emotions of persons is called Sentiment Analysis (SA) [1]. SA facilitates the task of finding words, phrases, and sentences that refer to sentiments and relieve to comprehend the relationship between textual reviews and the indication and values of these reviews [2].

The increased usage of websites and social networks allowed and enabled people to express their views, and with the existence of such vast amount of textual reviews and opinions, and according to [1], the need has increased to have a systemic model that has the ability to analyze such massive content and asserts it to (positive, negative, or neutral) opinions and then track attitudes in different domains such as political, governmental, and commercial fields. The existence of such huge amount and different types of information on the social media makes it very difficult to review, thus leading to a necessity to create new techniques and ways to deal with, and figure out, intentions behind opinions. Opinions are a focal activity to nearly all people because it is the way that people express their point of view about many things in our world, thus, enabling them to make decisions. This has generated an increasing attention in mining opinions [3]. One of the major tasks of mining views, or opinions, is to identify the polarity of such view or opinion.

General types of information that could be found on the Internet are objective and subjective. Target Information comes in a set of forms that can be observed, or facts, that are the closest to the truth. Personal information refers to judgments, opinions, beliefs, assumptions, and doubts that differ from person to person and from day to day. This type of information can cause disruption when someone has an option to make [4].

Research in the field of Arabic Natural Language Processing (ANLP) has lagged behind research in its counterpart Latin-based NLP due to many reasons to name some: the complexity of the language itself, the lack of support for Arabic by computers in the early history of computers, the

lack of standard machine-readable resources, as well as the lack, or hesitation, of researchers and/or users who are interested in working with Arabic.

As indicated by many researchers [6], [7]; work on SA has been in existence for the past one and a half decades but has targeted mostly Latin-based languages like English. Work on languages other than Latin-based languages has lagged and is still at its early stages. Even though SA in other languages has achieved great progress, many types of research, especially, in the field of Arabic SA have started to appear recently [8].

In these days, SA is considered as the main source to obtaining accurate information from a wide range of people without the need to ask them to fill direct surveys. Many organizations and companies focus nowadays on collecting and extracting users' opinions for different fields, especially for marketing and advertising as well as political reasons, to mainly understand the impact of such opinions on economics and public relations [6]. The need to deal with these unstructured and unregulated data naturally has led to an increased research into SA.

As for extremism, the term itself might vary based on the domain being inspected. For instance, if we are dealing with a political point of view or religious views, then extremism might refer to violent radicalization. Using this context, much research was conducted like that of [10], [11], [12], and [13]. Radicalism, as extremism, will not be covered in this research and is left to be a future research. In this research, however, we focus on another context of extremism, i.e. the task of finding outliers in different Arabic reviews and opinions and classifying them as extreme reviews and opinions by applying machine learning techniques.

Subsequently, this research area is being much substantial due to the fast growth of social media that enable all users to post their sentiments, comments, tweets, opinions, recommendations, reviews, and emotions [5].

1.1 Background

Since Arabic is the fourth spoken language and the biggest language from the Semitic dialect family and one of the fastest growing languages on the Internet, the number of Arab users has grown by a rate of 6.6% yearly and Arabic sentiment analysis was identified as an important research part in the field of SA [9]. Two types of Arabic languages are used nowadays: 1) Formal written Arabic: which consists of Modern Standard Arabic (MSA) as well as the Classical Arabic, and 2) Informal Arabic (day-to-day spoken Arabic): this type does not follow any grammatical rules or spelling standards like that of the formal types. Colloquial/Dialectal Arabic is a subtype of informal Arabic that exists based on the geographical area, or location in a country, and with the fact that sometimes the same country may have several dialects.

Since applying the SA techniques on all languages will be identifying the polarity of each sentiment to either positive, negative, or neutral in a one step process, we cannot use the same approach to identify any outliers, or extreme, opinions for a given text.

Since Sentiment Analysis techniques are used mainly to determine the polarity of a given text to one of three polarities (Positive, Neutral, and Negative), it will not be easy to use them to determine extreme positive or extreme negative polarities as this would require more efforts. For this, the main idea of this research is to divide the process into two stages. The first stage aims at classifying Arabic reviews into either positive or negative reviews by feeding the machine leaning algorithms with cleaned and processed reviews using some cleaning and NLP steps. In the second stage, the results of the first stage are then relabeled using ten rules that were extracted from a human majority questionnaire to classify positive and negative reviews, or opinions, further into positive, extreme positive, negative, and extreme negative reviews and then using the machine learning algorithms again on the produced dataset to identify extreme reviews properly.

The contribution of this paper is three-fold: first, it introduces a new approach for SA (Two-Stage) to find extreme reviews, second it proposed a new sentiment analysis dataset that was collected from Google play store, while third, it investigates three machine learning techniques to classify extreme reviews from Arabic opinions, a research that was not carried out before (according to the authors' knowledge).

2. RELATED WORK

This research is dealing with Sentiment Analysis for Arabic reviews and applying suitable machine learning techniques in a two-stage approach to find extreme reviews. For this, in this section we talk about research related to Arabic Dataset collection, Arabic NLP, Sentiment Analysis, and Machine learning.

Some research was conducted to describe dataset collection suitable for Arabic Sentiment Analysis, such as the works of. [14] who introduced an Arabic Corpus for Opinion mining named OCA that consist of 500 movie reviews that were collected from different web sites and blogs and was manually categorized as 250 positive and 250 negative reviews; and the work of [15] who proposed an approach to collect (2000) business reviews from Arabic websites for the purposes of Sentiment Analysis. In the work of [16], two Arabic Corpora; the Opinion Corpus for Arabic (OCA) and the Arabic Corpus

for Opinion Mining (ACOM) that was manually classified into positive, negative, and neutral datasets were used. A large-scale dataset of Arabic book reviews suitable for SA containing 63,000 book reviews that was rated on a scale from 1-5 was built by [17]. A dataset of (1725) reviews about products and services was collected from http://jeeran.com website and used by [18]. In the work of [19], they have collected 10,000 tweets to build what is known as the Arabic Social Sentiment Analysis dataset (ASTD) that was classified into four categories: objective, subjective positive, subjective negative, and subjective mixed.

According to authors, the first publicly available Arabic Sentiment Analysis Lexicon using a combination of existing resources like English SentiWordnet (ESWN), Arabic WordNet, and the Standard Arabic Morphological Analyzer was built by [20]. In their research, [21] used an approach to build non-English sentiment lexicons using unannotated corpus that proved, according to authors, itself on Arabic in comparison to other methods and lexicons.

A supervised classification using Machine Learning approach to SA was built by [22] and applied an in-house dataset of 2591 tweets and/or Facebook comments. The researchers of [23] used tweets written in modern standard Arabic (MSA) about terrorism and political events that occurred in the Arab countries and classified them manually. In [24], a Machine Learning model to evaluate Arabic tweets using two machine learning algorithms Naïve Bayes and Decision Tree was built, and around 8053 Arabic YouTube comments were collected and labeled manually by [9] and some volunteering graduate students.

Many of the previously mentioned research attempts applied machine learning classifiers like Support Vector Machine (SVM), Naïve Bayes (NB), K-means Nearest Neighbor (KNN), and Decision Trees (DT) for SA on their collected datasets. Most of those studies showed that SVM, as well as NB classifiers, showed the best classification results.

Another approach to Sentiment Analysis was conducted by [1], who used Rough-set theory to identify the polarity of Arabic tweets (positive, negative, or neutral) using the reduction concept of the set theory to identify the necessary features from different tweets to determine the rules needed for classification.

3. METHODOLOGY

In order to apply our approach and enable the finding of extreme opinions, the main research steps that were followed are:

1) Collecting reviews from Google play store using AppBot software to build the dataset, step
2) Cleaning the dataset and manually annotating 1500 reviews (750 positives, 750 negative),
3) Pre-processing of cleaned dataset (stopword removal, tokenization, POS tagging),
4) Identifying features used to build the feature vector from the annotated reviews, and
5) Developing and testing the most known classifiers used for sentiment analysis.

An overview of the followed algorithm by the methodology is given in Figure 1.

3.1 Data Set Collection

Due to the lack of available Arabic resources on the Internet; including datasets that can be used for benchmarking purposes, we had to build our own dataset through using Appbot; a web tool that can categorize each application review automatically into their polarity class. The collected dataset consists of 2500 Arabic reviews about fifteen different applications on the Google Play Store like Facebook, Snapchat, Open Sooq, etc.

***Input**: Arabic google play store reviews*

Begin:
1: Read reviews from csv file
2: Preprocessing the reviews
2.1: Remove punctuation marks and quotation
2.2: Tokenize the reviews
2.3: Remove stopword
2.4: Apply PoS Tagging
3: Feature Selection
3.1: Create vector model
3.2: Create feature vector for every review
3.3: Calculate the weight of every vector using TF-IDF
4: Classification process
4.1: Apply ML techniques on dataset
4.2: Check the classification if positive assign positive polarity then go to step 5
4.2: Check the classification if Negative assign negative polarity then go to step 5
5: Generate positive and negative report files
6: Apply extremism rules to positive and negative files.
7: Generate too positive file and too negative file
8: Calculate the accuracy of each ML classifiers
9: Generate final report

***Output:** Arabic extremism sentiment analysis polarity file*

Figure 1: Research Methodology Algorithm

In order to use the collected dataset properly, we initially had to perform an initial data cleaning process via removing duplicated and repeated reviews. For instance the word "جميل" was repeated 56 times; which may affect the accuracy of the approach. This process was conducted manually to reduce repeated reviews to appear only once, which resulted in a dataset of 1500 unique reviews that were classified equally into 750 positive and 750 negative reviews.

3.2 Data Cleaning

In this process, we filtered the extracted reviews manually before the pre-processing steps. This process also included removing noise; such as any non-textual data, non-Arabic text, symbols (*, !, $,#,&,~) and any unnecessary information found in the review that is not required for the classification

process, such as the reviewer's name and date. To deal with Emoticons, Smileys, and Emoji checks, we changed and replaced each emoji in the reviews with their corresponding class label, or their meaning in the language, by adding a single Arabic vocabulary against each emoji according to their polarities. An example of Emoticons, Smileys, and Emoji checks is given in Figure 2.

Part of the cleaning process was to deal with what is known as Arabizi format. This process was used manually when there are Arabic words in the reviews that are written using some Latin letters and numbers. For this type of text, we manually converted them into their equivalent Arabic words using Arabic characters. An example of an Arabizi review from our dataset converted into its equivalent Arabic words is shown in Table (1).

Part of the preparation steps of the dataset was to correct misspellings of Arabic reviews and remove any letters that do not belong to words. For this purpose, we manually used MS Word dictionary as a reference for correcting the misspelled words by selecting the most suitable suggested word by the dictionary. Table (2) gives an example of a misspelled review and its correction.

Sentiment Polarity	Emotion Classes	Emoticons and Emoji ideograms related
Positive	Happiness Laugh Grin	:-):) :D:o):] :3:c):>=|8)
Negative	Sadness	>:[:-(:(:-c :c :-< :< :-[:[:{
	Anger	:-| :@ >:(
	Cry	:'-(:'(
	Horror	D:< D: D8 D; D= DX v.v D-'
	Skeptical Annoyed	>:>:/:-/:-. :-:=/=:L=L:S>.<

Figure 2: Example Emoticons and Smileys.

Table 1: Example of Arabizi Reviews from our Dataset Converted to Arabic.

Arabizi format	Ajmal tatbe8 feelkonkolo 3jeb
Arabic equivalent	اجمل تطبيق في الكون كلو عجيب

Table 2: Example of Misspelling Review from our Dataset and its Correction to Arabic.

Review	Content
Misspelled review	خاتم القران جيمل مشاء الله واضافه قوايم الاذكار منتازه
Corrected review	ختمة القرأن جميلة ما شاء الله وإضافة قوائم الاذكار ممتازة

3.3 Natural Language Processing Steps

In the NLP processing steps three operations to process the dataset are used; Tokenization, POS-Tagging, and Stopwords removal as follows:

A. Tokenization:

One of the most significant steps in text classification since it allows us to deal with each word separately. It is the process of splitting statements, or strings, into words called tokens [25] (25). White spaces, commas, semicolons, colons, and dots can be used as a separator, or delimiters. These tokens can be single words (noun, adjective, adverb, verb, pronoun, etc.). The outcomes from the tokenization process are not totally meaningful as the tokenization process considers the "punctuation marks" as a token. For this, a second step is carried out to remove the punctuation. Table (3) gives an example of a review from our dataset and their tokens.

B. POS-Tagging:

Parses the document, or the statements, into words or terms and assign them into their corresponding part of speech (POS); a noun, verb, adverb, adjective, etc. In this step the Stanford part-of-speech tagger was used which supports different languages including Arabic. The tagger is implemented in many programming languages such as Java, PHP, C# and Python. This tagger takes the output from the tokenization process to determine the tag of each word. POS-Tagging helps in extracting the sentiment features; adjectives and adverbs can be identified as sentiment features. Table (4) gives an example of a review from our dataset and its POS tagging.

Table 3: Example of a Review from our Dataset and its Tokens.

Type	Content
Review	دائمًاعلي لسانك " ذكرالله يجعل رائع "تطبيق
Tokens	Token1:"تطبيق ", Token2:" رائع ", Token3:"يجعل", Token4:" ذكر ", Token5:" الله ",Token6:"دائمًا", Token7:"علي ", Token8:" لسانك "

C. Stopword Removal:

Stopwords are the most typical and frequent used words in a text that have no significant semantic meaning [25]. The stopwords removal process eliminates these words from the text (negation, pronouns, prepositions) that do not carry any polarity, or are not useful in the classification part; examples include prepositions (" في" (in), "من" (of), "على" (on)), demonstratives, ("هذا"(this), "هؤلاء")these (, and " اولئك"(Those)), interrogatives (" أين") where, ("متى") when ,("من") whom)) etc. There are several sources of stopword lists that are publicly available, but in this research, we used the Arabic stopwords of NLTK. The popular strategy for identifying a stopword list is to calculate the frequency of appearance of a word in a dataset then mark the most frequent words as stopwords list. Many conditions may be used to determine the stopword terms in a dataset, stopwords do not give any meaning if they appear alone in the text, used many times in the text, general terms or used in different fields.

Table 4: Example of a review from our dataset and its POS tagging.

Type	Content
Review	جربت هذه التطبيق وكانت تجربه حلوه ومفيده شكرا
Translation	I tried this app and it was a nice and useful experience
POS Tagging	('جربت/VBD'), ('هذه/NN'), ('التطبيق/DTNN'), ('الرائع/DTJJ'), ('وكانت/VBD'), ('تجربه/NN'), ('حلوه/NNP'), ('ومفيده/NNP'),(" شكرا/NN')

D. Text Transformation

After preprocessing of reviews, the remaining terms of each review is then transformed into a vector [26]. The vector is then used to calculate the weights of the included terms using the Term Frequency-Inverse Document Frequency (TF-IDF) measure; a technique to assess how much the words are important in a document. TF-IDF is also used to convert the text into vectors, or a vector space model, to find features in a document or a text. The vector space model is away, or a method, used to represent the text as a vector, which represents important words or features, or the presence, or the absence, of terms Bag of Words (BOW) in the document. The following equations are used to calculate TF-IDF values for each word in a document:

$$TF(i,j) = \frac{Frequancy\ of\ term\ t\ in\ review\ j}{Total\ number\ of\ terms\ in\ review j} \quad (1)$$

$$IDF(i,j) = log \frac{Total\ \#\ of\ reviews\ in\ the\ dataset}{\#\ of\ reviews\ which\ include\ term\ i} \quad (2)$$

$$TF - IDF = TF(i,j) * IDF(i,j) \quad (3)$$

4. EXPERIMENTS

In our study, we focused on Arabic sentiment analysis at the document-level. However, our main used approach is a machine learning approach. Since employing Machine Learning for SA is not directly applicable in finding extreme reviews from raw data, we followed a classification process that uses two stages.

In the first stage, we employ ML classifiers with different X-fold values for splitting the dataset to find the sentiment polarity of text reviews and measuring the accuracy of each classifier. The resulting output from the first stage, is then fed as an input to the second classification stage; a stage aimed to identify the extreme reviews from the output of the first classification stage.

To accurately classify opinions in this stage, some rules were also written for the Arabic opinions and reviews that aim to add more weight to the words of a document; such as intensification words in Arabic like (كثيراً, جداً, بافراط, بزياده) which makes any review containing one of these words an extreme review.

The Second stage classification is highly dependent on human annotations, as the extreme rules were extracted from answers to the questionnaires by native Arabic speakers to identify extreme reviews from normal ones.

Ten different rules were built after analyzing the questionnaire and annotator's feedback. If any review satisfies one, or more, of these rules, then it is considered as an extreme review and the review is then used to build the extreme dataset. According to the approach, the evaluation of the sentiment classification depends on the annotated extreme dataset, which was built for the purposes of classification in the second stage. The outputs of the sentiment classification experiments are presented as follows:

4.1 First Stage Classification:

In this stage, two main Document-Level SA experiments are conducted using Different ML Classifiers (NB, SVM, RF) with TF-IDF feature selection and different Cross Validation schemas. In the first experiment the dataset was split into (30-70), while in the second experiment, the dataset was split into (20-80) for testing and training respectively.

4.1.1 Experiment A

In this experiment, we conducted a Document-Level Sentiment Analysis using Different ML Classifiers with TF-IDF Feature Selection and Cross Validation (30-70). The dataset was split into training data and testing data according to cross-validation, where 1500 reviews were split into 30% (450) reviews for testing the model and 70% (1050) reviews for training the model. This experiment included three sub-experiments and their results are given in Tables 5, 6, and 7.

Table 5: Performance results of NB classifiers with 30% Testing data and TF-IDF feature.

	Precision	Recall	F-score	Support
Negative	0.91	0.77	0.84	222
Positive	0.81	0.93	0.86	228
Total / average	0.86	0.85	0.85	450

Table 6: Performance results of SVM classifiers with 30% Testing data and TF-IDF feature.

	Precision	Recall	F-score	Support
Negative	0.83	0.95	0.88	222
Positive	0.94	0.81	0.87	228
Total / average	0.89	0.88	0.88	450

Table 7: Performance results of RF classifiers with 30% Testing data and TF-IDF feature.

	Precision	Recall	F-score	Support
Negative	0.77	0.94	0.85	222
Positive	0.93	0.72	0.81	228
Total / average	0.85	0.83	0.83	450

4.1.2 Result Discussion of Experiment A

By comparing the results of the three machine learning classifiers from tables 5, 6, and 7, the overall precision of the NB is 86% with a higher value of negative precision of 91% and positive precision of 81%. However, the recall of the negative predictions is 77% and is lower than the positive predictions of 93%, which indicates that the system has misclassified more negative reviews than positive ones. Since the F-measure depends on the values of recall and precision for the classifiers; it will follow both previous measurers, NB has a total F-score value of 85%.

On the other hand, the Random Forest classifier showed an overall precision of 85% with a higher positive precision of 93% and a lower negative precision of 77%. However, the recall of the positive predications is 72% and is lower than the negative predications of 94%, which indicates that the system has misclassified more positive reviews than negative ones with 83% total F-score.

SVM classifier, however, showed the best recall and precision results compared with other classifiers; it showed an overall precision of 89% with higher positive precision of 94% and 83% of negative precision, while the negative recall has the best result ever with a 95% values; which means that the SVM didn't misclassified too much negative reviews, and an 81% positive recall with a total recall of 88% and a total F-score of 88%. Higher recall values mean fewer false negative reviews; positive reviews which have been wrongfully classified by the classifier as negative, while lower recall values mean more false negative reviews. Improving recall values often decreases precision. However, higher precision means fewer false positive; negative reviews which have been wrongfully classified as positive, while lower precision values means more false positive predictions.

4.1.3 Experiment B

In this experiment, we conducted a Document-Level Sentiment Analysis using Different ML Classifiers with TF-IDF Feature Selection and Cross Validation of (20-80). The dataset of 1500 reviews were split into 20% (300) reviews for testing the model and 80% (1200) reviews for training the model. This experiment included three sub-experiments as well. Tables 8, 9, and 10 give the results for such experiments.

Table 8: Performance results of NB classifier with 20% Testing data and TF-IDF feature

	Precision	Recall	F-score	Support
Negative	0.90	0.81	0.85	139
Positive	0.85	0.92	0.88	161
Total / average	0.87	0.87	0.87	300

4.1.4 Result Discussion of Experiment B

Comparing the results of the three-machine learning classifiers from tables 8, 9, and 10, an overall precision of 87% was yield by SVM with a lower value of negative precision of 79% and positive precision of 93%. However, the recall of the negative predictions of 94% is higher than the positive predictions of 78%, which indicates that the system has misclassified more positive reviews than negative ones. Since the F-Measure depends on the values of recall and precision for the classifiers, SVM has a total F-score value of 85%. While the Random forest classifier showed an overall precision of 77% with a higher positive precision of 78% and a lower negative precision of 76%. However, the recall of the negative predictions of 74% is lower than the positive predictions of 80%, which indicates that the system has misclassified more negative reviews than positive ones with 77% F-score. NB classifier showed the best recall and precision results compared to other classifiers; it showed an overall precision of 87% with higher negative precision of 90% and 85% of positive precision, while a positive recall of 92%; which means that the NB didn't misclassified too much positive reviews, with 81% negative recall and a total recall of 87% and a total F-score of 87%.

Table 9: Performance results of SVM classifier with 20% Testing data and TF-IDF feature

	Precision	Recall	F-score	Support
Negative	0.79	0.94	0.86	139
Positive	0.93	0.78	0.85	161
Total / average	0.87	0.85	0.85	300

Table 10: Performance results of RF classifier with 20% Testing data and TF-IDF feature

	Precision	Recall	F-score	Support
Negative	0.76	0.74	0.75	139
Positive	0.78	0.80	0.79	161
Total / average	0.77	0.77	0.77	300

4.1.5 Result Discussion for Both Experiments A & B

In order to choose the best classifier for this stage properly, we have calculated the accuracy (A) of the results using the

following formula, where TP refers to True Positive, TN refers to True Negative, FP refers to False Positive, and FN refers to false Negative. Table 11 shows a summary accuracy results obtained during the first stage classification by the three classifiers: SVM, NB, and RF using TF-IDF scheme with both 30% and 20% test data.

$$A = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

By comparing the results of the used classifiers, and when using a 30% test data, it is clear that SVM has better results than the other two classifiers with a total accuracy of 87.7% while NB accuracy was 85.1% and RF showed the lowest accuracy of 78.6%. This result was also observed in more than one study, like that of [4]. We can also notice from the same table that using the NB with a TF-IDF and 20% for test data gives the best accuracy with 86.6%, while SVM accuracy was 85.3%, and RF showed an accuracy of 78.6%. The obtained results from the first stage were used as an input to the second stage classification to identify extremism of Arabic opinions.

Table 11: Total Accuracy for the Three ML Classifiers with Different Cross-validation scheme

	SVM	NB	RF
Total accuracy with 30% test data	0.877	0.851	0.786
Total accuracy with 20% test data	0.853	0.866	0.786

From the accuracy results in table 11, we decided to use SVM with 30% testing data and NB with 20% testing data to find extreme positive reviews from the produced predictions from NB classifier for positive reviews and the same thing to find extreme negative reviews from the produced negative predictions using SVM classifier with 30% testing data and NB with 20% testing data.

Since the produced data from both classifiers is not balanced, as the number of produced negative reviews from the SVM with 30% testing data is 222 and the positive reviews are 228, and the number of produced negative reviews from the NB with 20% testing data is 139 and the positive reviews is 161 and the total number of predicted positive reviews from the two classifiers is 389 and the number of predicted negative reviews is 361 totaling 750 reviews, we decided to build a new balanced and accurate dataset from the produced predictions for both classifiers. Since the classifier's accuracy measurement is not 100% correct, human annotation has played a major role in labeling the reviews into their correct corresponding class, the new and final produced dataset, after balancing and ensuring the correct labeling of each review to be used in the second stage classification, we ended up with a balanced dataset of 700 reviews (350 positive reviews and 350 negative reviews).

4.2 Second Stage Classification

This phase is aimed to find extreme reviews from all reviews of the new dataset. Since SA is a binary classification problem, the built dataset was split into 350 confirmed positive and 350 confirmed negative reviews. In order to apply the SA to the positive and negative datasets and to find the extreme reviews from the normal ones, the positive dataset was also annotated into positive and too positive reviews by distributing a questionnaire to a group of graduate students at the Information Systems Department of Yarmouk University to identify extreme reviews based on some handwritten guidelines and the built rules; such as the appearance of the word (جدا), the existence of repeated letters in the same word such as (بجنننننن), and in addition to the usage of some colloquial words that reflect the overall meaning of the review such as (طقع). The positive dataset consists of 350 reviews which were labeled into 175 positive reviews and 175 extreme (too positive) reviews. In this stage, two main Document-Level SA experiments are conducted using Different ML Classifiers (NB, SVM) with TF-IDF Feature Selection and different Cross Validation schemas. In the first experiment the dataset was split into (30-70), while in the second experiment the split was (20-80) for testing and training.

4.2.1 Experiment C:

In this experiment, we conducted a Document-Level Sentiment Analysis using Different ML Classifiers with TF-IDF Feature Selection and different Cross Validation schemes to find extreme positive reviews from the positive dataset. Using SVM, the dataset of 350 reviews was split into 30% (105) reviews for testing and 70% (245) reviews for training the model. As for the NB, we used 20% (70) reviews for testing and 80% (280) reviews for training the model. The results of this experiment are shown in Tables 12 and 13.

Table 12: Performance results of SVM (with 30% testing) Sentiment Classification on Positive reviews dataset.

	Precision	Recall	F-score	Support
Positive	0.65	0.78	0.71	51
Too Positive	0.74	0.59	0.66	54
Total / average	0.70	0.69	0.68	105

Table 13: Performance results of NB (with 20% testing) Sentiment Classification on Positive reviews dataset.

	Precision	Recall	F-score	Support
Positive	0.79	0.42	0.55	36
Too Positive	0.59	0.88	0.71	34
Total / average	0.69	0.64	0.62	70

4.2.2 Result Discussion of Experiment C

By comparing the results of the two machine learning classifiers from Tables 12 and 13, the overall precision of NB was 69% with a higher value of positive precision of 79% and too- positive precision of 59%. However, the recall of the positive predictions of 42% is much lower than the too-positive predications of 88%, which indicates that the system has misclassified more positive reviews than too-positive ones, NB has a total F-score value of 62%. The SVM classifier showed the best recall and precision results compared to the NB classifier; it showed an overall precision of 70% with higher too-positive precision of 74% and 65% for positive precision respectively, while the positive recall has the best results; ever with 78%. This means that SVM didn't misclassified too much positive reviews, and a recall of 59% for too-positive with a total recall of 69% and a total F-score of 68%.

4.2.3 Experiment D

In this experiment, we conducted a Document-Level Sentiment Analysis using Different ML Classifiers with TF-IDF Feature Selection and with different Cross Validation scheme to find extreme negative reviews from the negative dataset. The dataset was split into training and testing data according to cross-validation. SVM using the dataset of 350 reviews was split into 30% (105) reviews for testing the model and 70% (245) reviews for training the model. as for the NB we used 20% (70) reviews for testing the model and 80%(280) reviews for training the model.

Table 14: Performance results of SVM (with 30% testing) Sentiment Classification on Negative reviews dataset.

	Precision	Recall	F-score	Support
Negative	0.62	0.76	0.68	51
Too Negative	0.71	0.56	0.63	54
Total / average	0.67	0.66	0.65	105

Table 15: Performance results of NB (with 20% testing) Sentiment Classification on Negative reviews dataset.

	Precision	Recall	F-score	Support
Negative	0.81	0.61	0.70	36
Too Negative	0.67	0.85	0.75	34
Total / average	0.75	0.73	0.73	70

4.2.4 Result Discussion of Experiment D

By comparing the results of the two machine learning classifiers in Tables 14 and 15, the overall precision of NB was 75% with a higher value of negative precision of 81% and too-negative precision of 67%. However, the recall of the negative predictions of 61% is lower than that of the too-negative predictions of 85%, which indicates that the system has misclassified more too-negative reviews than negative ones. NB has obtained a total F-score value of 73%. The SVM classifier showed an overall precision of 67% with higher too-negative precision of 71% and 62% of negative precision, while the negative recall result with 76% and 56% too-negative recall with a total recall of 66%; which means that the SVM didn't misclassified too much negative reviews with a total F-score of 65%.

4.2.5 Result Discussion of C & D Experiments

In order to choose the best classifier for this stage properly, we have calculated the accuracy (A) using formula 3, where TP, TN, FP, and FN refer to the same concepts. Table 16 shows the accuracy results obtained in the second stage of classification using both classifiers; NB; with 20% for testing data, and SVM; with 30% for testing data, both using TF-IDF scheme. By comparing the results of the two classifiers, it is clear that SVM has scored better results when dealing with the positive dataset with an overall accuracy of 68.7% with 65.7% accuracy for the negative dataset. The NB, on the other hand, showed better accuracy when dealing with the negative dataset, with an overall accuracy of 72.8%, with an accuracy of 64.2% for the positive dataset.

Table 16: Total Accuracy results for the Two ML Classifiers with Different Cross-validation schemes.

	SVM with 30% Test Data	NB with 20% Test data
Total accuracy for the negative dataset	0.657	0.728
Total accuracy for the positive dataset	0.687	0.642

5. CONCLUSIONS

SA techniques are already being used through many applications to determine the polarity of information (Positive, Negative, or Neutral). However, a normal (One-Stage) Sentiment Analysis cannot be used easily to identify extreme and outlier opinions. The sentiment of different Google Play Store applications' reviews was investigated, and a combination of different pre-processing methods was employed to reduce the noise in the text in addition to using the TF-IDF method to determine the important features that affect the text polarity.

To identify extremism in application reviews; we performed a two-stage classification process to our dataset. To test the validity of this approach, machine learning classifiers like SVM, NB and RF were used to classify the sentiment of reviews. The predicted and refined reviews from the first stage were used to perform a second classification phase to find extreme reviews. As a result, it was found out that SVM has outperformed the NB classifier.

The first stage used three different machine learning classifiers with different cross-validation values and TF-IDF feature model, the results showed that using SVM with 30% testing data has achieved higher accuracy results of 87.7%, while NB with 20% test data has achieved a total accuracy result of 86.6%. With RF scoring the worst results among all classifiers, it was ignored for the second stage. The prediction results from both adopted classifiers lead to the produced new dataset, where different rules were applied to the produced dataset to label it into four classes (positive, too positive, negative, too negative) then the second classification stage was applied, where the new dataset was split into a positive dataset; containing positive and too positive reviews, and a negative dataset; containing negative and too negative reviews.

To investigate the performance of the classifiers, the same accuracy metrics were applied for the second classification stage; which showed good results; where NB with 20% test data showed better accuracy results with the negative dataset of 72.8 %, but the SVM with 30% showed better accuracy results than NB with 20% test data for the positive dataset with a total accuracy of 68.7%.

6. FUTURE WORK

There are many challenges that faced us in building our approach, in particularly the lack of Arabic resources that support our approach, especially, to address Arabic extreme content. Some of these challenges and problems were resolved in our approach while others remained and determined unresolved. For future work, we will continue in the line of research. The limitation of our approach lies in the following which constitutes some possible direction of future work:

- Employ our approach to larger datasets; especially for the second classification phase.
- Improve the dataset itself to include rating stars with the reviews.
- Employ other machine learning classifiers such as K-Nearest Neighbors (KNN).
- Expand the approach to deal with different deep learning classifiers.
- Try different feature selection methods, informatics features, or semantic features such as Part of speech (POS), N-gram, and negation.
- Create a hybrid approach which combines both Machine Learning (ML) and Semantic Orientation (SO) approach.

REFERENCES

1. Al-Radaideh, Q. and Al-Qudah G.**"Application of Rough Set-Based Feature Selection for Arabic Sentiment Analysis"**. *Cognitive Computation*, Vol. 9, no. 4, pp. 436-445, 2017.
2. Thakare, A.N., & Wankhede, R.. **Design approach for accuracy in movies reviews using sentiment analysis**, in *Proc. International conference of Electronics, Communication and Aerospace Technology (ICECA)*, 2017, 1, pp. 6-11.
3. Albraheem L. and Al-Khalifa H. **Exploring the problems of sentiment analysis in informal Arabic**. In *Proc. of the 14th International Conference on Information Integration and Web-based Applications & Services*, 2012, pp.415-418.
4. Saleh M. **Sentiment analysis in Arabic: Opinion Polarity Detection.** PhD dissertation, UNIVERSIDAD DE JAĖN, 2013.
5. Baca-Gomez Y., Martinez A., Rosso P., Esquivel H., and Hernandez-Farias D. **Web service SWePT: A hybrid opinion mining approach.** *Journal of Universal Computer Science*. Vol. 22, no. 5, pp. 671-690, 2016.
6. Ravi K., and Ravi V. **A survey on opinion mining and sentiment analysis: tasks, approaches and applications**. *Knowledge-Based Systems*, Vol. 89, pp. 14-46, 2015.
7. Qin Z.. **A framework and practical implementation for sentiment analysis and aspect exploration**. Unpublished PhD. dissertation, University of Manchester, 2017.
8. Baly R., Badaro G., El-Khoury G., Moukalled R., Aoun R., Hajj H., and Shaban K. **A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models**. In *Proc. of the Third Arabic Natural Language Processing Workshop*, pp. 110-118, 2017.

9. Al-Tamimi, A. K., Shatnawi, A., & Bani-Issa, E.. **Arabic sentiment analysis of YouTube comments.** In *Proc. Of Applied Electrical Engineering and Computing Technologies (AEECT), IEEE Jordan Conference on,* 2017 pp. 1-6.
10. Bermingham A., Conway M., McInerney L., O'Hare N., and Smeaton A. **Combining social network analysis and sentiment analysis to explore the potential for online radicalization**. In *Proc. of the International Conference on Advances in Social Network Analysis and Mining (ASONAM'09)*, 2009, pp. 231-236.
11. Chalothorn T., and Ellman J. **Using SentiWordNet and sentiment analysis for detecting radical content on web forums**. In *Proc. of the 6th Conference on Software, Knowledge, Information Management and Applications (SKIMA 2012)*, pp. 9-11.
12. Iskandar B. **Terrorism Detection Based on Sentiment Analysis Using Machine Learning**. *Journal of Engineering and Applied Sciences*, Vol. 12, no. 3, pp. 691-698, 2017.
13. Kocharekar M. and Jadhav U. **Detecting Terrorist Activities using Sentiment Analysis**. *Distributed Systems*. Vol. 6, no. 3, pp. 285–287, 2017.
14. Saleh R, Martín-Valdivia M.,Ureña-López L., and Perea-Ortega, J. **OCA: Opinion corpus for Arabic**. *Journal of the American Society for Information Science and Technology*, Vol. 62, no. 10, pp. 2045-2054, 2011.
15. Elhawary M. and Elfeky M. **Mining Arabic business reviews**. In *Proc. of Data Mining Workshops (ICDMW), IEEE International Conference on, 2010*, pp. 1108-1113.
16. Mountassir H., Benbrahim I., and Berrada. I. **A crossstudy of Sentiment Classification on Arabic corpora**. *Research and Development in Intelligent Systems XXIX*, Springer, pp. 259–272, 2012.
17. Aly M. and Atiya A. **Labr: A large scale arabic book reviews dataset**. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics. 2*, 2013, pp. 494-498.
18. Omar N., Albared M., Al-Shabi A., and Al-Moslmi T. **Ensemble of classification algorithms for subjectivity and sentiment analysis of Arabic customers' reviews.** *International Journal of Advancements in Computing Technology (IJACT)*, Vol. 5, no. 14, pp. 77-85. 2013.
19. Nabil M., Aly M., and Atiya A. **ASTD: Arabic Sentiment Tweets Dataset**. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519.
20. Badaro G., Baly R., Hajj H., Habashand N. El-Hajj W. **A large scale Arabic sentiment lexicon for Arabic opinion mining**. In *Proc. of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 2014, pp. 165-173.
21. Mohammed K. and Balakrishnan V. **An automatic non-English sentiment lexicon builder using unannotated corpus**. *Journal of Supercomputing*, Vol. 75, pp. 2243–2268, 2019.
22. Duwairi R. and Qarqaz I. **Arabic sentiment analysis using supervised classification**. In *Proc. of Future Internet of Things and Cloud (FiCloud), IEEE International Conference on*, 2014, pp. 579-583.
23. Bouchlaghem R., Elkhelifi A., and Faiz R. **SVM based approach for opinion classification in Arabic written tweets.** In *Proc. of Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th IEEE International Conference of*, 2015, pp. 1-4.
24. Al-Horaibi L. and Khan M. **Sentiment analysis of Arabic tweets using text mining techniques**. In *Proc. of First International Workshop on Pattern Recognition*, 10011, 100111F), 2016.
25. Shoukry A. and Rafea A. **Sentence-level Arabic sentiment analysis**. In *Proc. of Collaboration Technologies and Systems (CTS)*, IEEE International Conference on, 2012, pp. 546-550.
26. Tubishat M., Abushariah M., Idris N., and Aljarah I.. **Improved whale optimization algorithm for feature selection in Arabic sentiment analysis**. *Applied Intelligence*, Vol. 49, pp. 1688-1707, 2019.