IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 19, 2019, accepted January 1, 2020, date of publication January 9, 2020, date of current version January 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2965181

Startup Initiative Response Analysis (SIRA) Framework for Analyzing Startup Initiatives on Twitter

BASHAYER ALOTAIBI[1], RABEEH AYAZ ABBASI[2], MUHAMMAD AHTISHAM ASLAM[1], KAWTHER SAEEDI[1], AND DIMAH ALAHMADI[1]

[1]Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, Quaid-i-Azam University, Islamabad 45320, Pakistan

Corresponding author: Rabeeh Ayaz Abbasi (rabbasi@qau.edu.pk)

ABSTRACT Social Media (SM) platforms, particularly Twitter, have become useful tools for startup companies (henceforth startups) which use the latter to support most of their business activities. As a result, there is a need to gauge the performance of specific business initiatives vis-à-vis public sentiment, or more specifically the spread of such initiatives based on Twitter user-generated content. Previous research which makes use of Twitter analysis to analyze the business activities of startups is minimal, especially for Twitter user content in the Arabic language. Consequently, this paper proposes an analytics-based framework called Startup Initiatives Response Analysis (SIRA) designed to assess the performance of initiatives launched by startups via text classification, sentiment analysis, and statistical analysis techniques. To provide empirical evidence for the viability of the proposed research framework, this paper examined the case of an Arab transportation network startup, carrying out a SIRA analysis of an initiative undertaken by Careem to empower women by encouraging them to work for the company. The results confirm the effectiveness of the proposed framework for statistically measuring the initiative spread and the public feedback based on the user-generated content on the Twitter social platform.

INDEX TERMS Data mining, machine learning, sentiment analysis, startups, entrepreneurship, Twitter.

I. INTRODUCTION

Social Media Analytics (SMA) has recently emerged as an essential approach for collecting and analyzing data from social media platforms. It uses advanced analytics tools and techniques to collect, process, and analyze Social Media (SM) data in order to identify useful patterns and knowledge [1]. SMA has been applied across a broad range of industries, including healthcare [2], social science [3], political science [4] and economy and business [5], [6] with a view to generating useful patterns that support various applications and activities.

With increasing use of SM platforms by companies, the practice has become an essential part of many business strategies. In particular, startups depend on such platforms to establish a strong business presence and maintain robust growth. To clarify the concept of startups in business domain, although it is stated that there is no clear definition of startup in terms of what they do [7], the NESTA definition cited in [7] defines startups as ''young, innovative, growth-oriented business (employees, revenue, customers) in search of a sustainable and scalable business model''. Hence, any business irrespective of its nature can be classified as a startup if it aims to develop a product or service that has not been offered before, demonstrates innovativeness in seeking to fill a gap in the market or changes the traditional way of doing something and focuses on business growth.

Microblogging platforms play a crucial role in enabling startups to overcome their limited resources and reach out to a wider audience, to perform effective business operations and to increase customer satisfaction. Furthermore, microblogging is a useful electronic word-of-mouth (e-WOM) tool through which consumers are able to share opinions. These opinions are expressed in blog post form or as written comments that enhance businesses by building effective relationships between businesses and customers [8], [9]. In recent

The associate editor coordinating the review of this manuscript and approving it for publication was Zhe Xiao.

years, a great number of companies have been interested in the power of SMA to support their decision-making processes, create new value and enhance their competitive advantage [1], [10]. Research efforts in analyzing startups activities on SM, specifically Twitter, via SMA are generally limited. Extant research which has been limited to European and American startups (i.e., English and Europe languages tweets) has focused largely on predicting success of startups based on their SM activities, their financial success, or on examining public opinion regarding startups in general. Therefore, this study seeks to fill the gaps in existing research by making the following contributions to literature:

1) Propose a novel Twitter-based analytics framework that enables startups to assess the performance of specific initiatives based on a comprehensive quantitative (statistical analysis) and qualitative analysis (text mining and sentiment analysis) techniques.
2) Choose an Arabic startup firm as a case study, so as to furnish empirical evidence for demonstrating the validity of the proposed framework.
3) Undertake comparative analysis based on evaluation of the performance of numerous machine learning classifier algorithms.
4) Undertake a comparative analysis based on evaluating different combinations of preprocessing techniques to improve the classification performance of Arabic text.

The sections of the paper are organized as follows: Section II discusses related research, and Section III discusses the research problem, whereas Section IV presents the proposed framework. The experiment setup and experiment results analysis are presented in Sections V and VI. Finally, the conclusion of the study is presented in Section VII.

II. RELATED WORK

This section reviews the current state of research on Twitter analysis, focusing on Twitter activities undertaken by startup companies. The main aim of the review is to identify a gap in the selected field wherein additional research is merited. Accordingly, the review was guided by the following research question:

RQ: How are Twitter analysis techniques used to generate deep business insights for startups?

The approach to synthesizing the literature for this study comprises analysis and comparison of all the surveyed studies and subsuming of the relevant studies' aims and methodology under three main themes elaborated in the following sections.

A. PREDICTING STARTUP SUCCESS

Numerous studies aimed at predicting the success of startups from early development stages to various phases in the life of a startup have been conducted. The applied techniques were found to include machine learning, data mining techniques or Social Network analysis techniques and to be based on structured data, e.g. [11]–[13].

While one study [11] sought to investigate the usage of digital traces in predicting the early stage of startup survival, another investigation [12] studied the success of startups at early development stages. In the first study [11] discussed above, different machine learning algorithms were applied to predict whether a startup survived or failed, leading to the conclusion that with a context-specific text mining approach, 5-year survival predictions of diverse survival rates ranging from 50% to 10%, with an accuracy of up to 91% could be made. In contrast, the second study mentioned earlier [12] examined the possibility of using web-based open sources in predicting a startup's success rather than mining structured data sources. The dataset was based on Crunchbase database along with crawling company's web-based and LinkedIn people profiles. The evaluation metrics of prediction performance included the use of ROC-AUC classification metric and analysis of the Precision-Recall curve. The results of predictive model results were presented comprehensively, whereby the results were discussed in terms of answers to five sub-problems answers or as hypothesis evidence. Notably, this study [12] suggests some future improvements, such as use of content analysis techniques (e.g., sentiment analysis). In a similar manner, the work presented in another study [13] presents a predictive model for startup success based on consideration of several key factors at play during various phases in the life of a startup. This study proposes a system to predict the failure or success of startups on the basis of analysis of a dataset of 11,000 startups from the Crunchbase resource. The system was based on using data mining classifiers (e.g., Naïve Bayes, logistic regression, decision trees, etc.) for the predictive model. In conclusion, there are several key factors which significantly change the predictive models such as seed funding amount raised by the startup or the rounds of funding it goes through. Furthermore, two significant key factors such as the Burn Rate of the company and few severity factors also affect startup outcomes.

B. THE ROLE OF SOCIAL MEDIA USAGE IN STARTUP FINANCIAL SUCCESS

The electronic word-of-mouth (eWOM) platforms have a significant economic impact on entrepreneurs and startups. Some recent research [7], [14]–[16] has been conducted to investigate the role of Twitter usage among startups and corresponding financial success.

In one study [7], the purpose of the investigation was to address the lack of empirical research examining the correlation among Twitter usage in European Union (EU) startups with the total investment in startups per country. The founder data was extracted from F6S.com databases and the focus was on founders' personal Twitter accounts. The number of posts and followers was extracted, and the results suggested a statistically positive correlation between total startup founder presence and the total investment per country. In a similar effort [14], the role of SM in startup outcomes was examined by studying the activities of entrepreneurs which influenced the outcomes of their startups. The empirical analysis was

based on Twitter data. Eight machine learning methods were employed in a data preparation procedure to provide convincing evidence. The results confirmed that differences in entrepreneurs' Twitter presence and activities have a significant effect on further engagement and venture financing. In terms of Twitter content analysis, another study [16] aimed to use entrepreneurs' social media behavior as the basis for predicting the success of their crowdfunding activities. The entrepreneurs' data was extracted from Kickstart, and social network analysis and sentiment analysis techniques were adopted to analyze the content of entrepreneurs' Twitter posts. The results indicated that innovative startups which use Twitter effectively receive a big crowdfunding amount in total. Contrastingly, in another study [15] considered in the review, the aim of the research was to investigate how entrepreneurs' emotions are affected by the funding process which was done by comparing entrepreneurs' emotional ratios with the verification of the founding process. The hypothesis stated that ''the positive emotions increase when there is funding''. To verify that, two main financial databases were adopted, and Twitter text analysis was applied based on entrepreneurs' accounts. The positive and negative emotion ratios were calculated, and Stata software was used to analyze the results and to run the regression. The results confirmed the hypotheses.

C. EXAMINING PUBLIC OPINION AND USER-GENERATED CONTENT REGARDING STARTUP SERVICES

Some research has recently been conducted that examines sentiment, perception, preferences and opinions of consumers relevant to startup firms based on SM user-generated content, for instance [17]–[19].

In the first instance, the study [17] aimed to mathematically measure consumer trust sentiments towards traditional business models (B2C) and compare these to modern business model (C2C) marketplaces. The sentiment analysis was performed using SocialMention tool, and the taxi industry sector was chosen as a case study. The results showed that the percentage of negative sentiments towards modern taxi companies was higher than towards traditional taxis. Also, amongst modern taxi companies, Uber drew the highest percentage of negative perceptions regarding unexpected behavior of drivers along with lack of experience. In contrast, another investigation [18] explored whether the startup collaborative consumption platform (Uber) was perceived by consumers as a technological innovation or as an institutional disruption. This study [18] applied a tool that collected user-generated content based on specific keywords of Swedish language. According to the findings, Uber was perceived by consumers as comprising both a technological innovation and an institutional transformation, with the latter view being more dominant. Another study [19], proposed a novel framework for the development of new products and services that aim to support companies in decision-making processes by considering consumer opinions and sentiments. A use case was developed over the Uber App to generate evidence with regard to methodology performance. The protest effects around Uber in Portugal on consumer perception were analyzed, and the results were clearly stated. Besides the usefulness of the framework in evaluating company products or services, it contributes to creating an understanding of the competitor environment.

III. RESEARCH QUESTIONS

SM analytics techniques and text mining contribute to the generation of meaningful, in-depth and reliable business insights that can enhance the effectiveness of business operations. Most of the prior SM analysis research on startups has focused either on predicting a startup's success, examining their financial success, or on general consumer perception regarding startup image and their products or services in general. To date, there is an identifiable gap in research that combines examination of customer sentiments and the performance of a specific business activity (e.g., initiative, marketing campaign) through a general tweet analysis aspect. Additionally, all prior studies have been limited to startups in Europe and the U.S (i.e., dealing with English and European Languages). In view of the above, this study will help to address the previously mentioned research gaps based on Twitter analysis techniques. To fulfill this objective, this study proposes a framework for measuring the performance of a specific initiative taken by a startup firm in an Arab country. This framework contributes to delivering valuable insights regarding the performance of an initiative that was launched on Twitter by combining qualitative and quantitative tweet analyses. These insights might assist startup founders or entrepreneurs with measuring the efficiency of their business operations and supporting their decision-making process. Such an analysis framework must address the following questions:

- How do customers feel about such an initiative?
- What is the impact of such an initiative on key engagement features of tweets?
- How quickly did the initiative spread across people?

IV. THE PROPOSED FRAMEWORK

As discussed in the previous section, the main aim of this work is to propose a framework that investigates the use of Twitter data in measuring the performance of a specific business aspect (e.g., initiative, marketing campaign, etc.). Figure 1 shows the general framework architecture. Startup Initiatives Response Analysis (SIRA) is an analytic Twitter-based framework that is designed to support startups or entrepreneurs in improving the performance of their initiatives. It measures Twitter activities, customer satisfaction, and temporal spread. SIRA framework consists of three major components that must be performed sequentially. The following sub-sections explain the framework phases in detail:

A. DATA COLLECTION AND ANNOTATION PHASE

The Twitter platform represents a valuable source for data collection which is commonly used by data science and

social network analysis researchers. The Twitter platform has become the preferred source for data collection due to the availability of free Twitter crawling tools. It is open source, which enables the developer to build a crawler script that queries and fetches tweets with all tweet metadata or descriptive information. There are numerous Twitter crawling open source libraries and tools that use Python or Java as base language. For example, Twitter API is widely used for tweet crawling; it is a public platform for querying public streams of information.

In this framework, the data is crawled using keywords, hashtags, Twitter accounts and other query parameters that are predefined by the user to search through Twitter. The extracted dataset is used for classification aim which is carried out using a supervised ML approach. Based on this, a training dataset is required that can be annotated either manually by a human or, in the case of a huge dataset, through crowdsourcing or through Unsupervised ML algorithms (i.e., lexical-based classifier). Text annotation is the process of adding a label or tag to the text based on pre-identified classes, such as the two classes 'others' and 'relevance' shown in Figure 1.

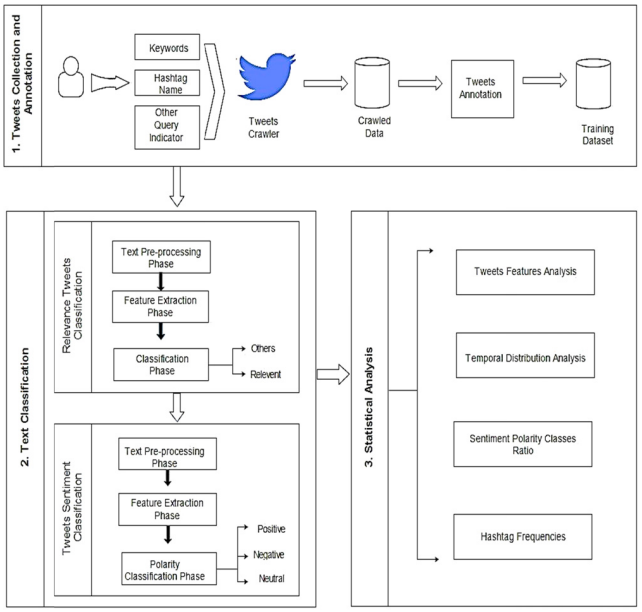

FIGURE 1. The proposed SIRA framework.

B. TEXT CLASSIFICATION PHASE

In the proposed (SIRA) framework, two text classifiers are developed. The first one, Relevance Tweets Classifier is a binary classifier that classifies each tweet as either relevant to the initiative subject or not (i.e., relevant or others). The first classifier is built because even with crawling based on specified keywords, there are some anomalous tweets. The second one, Tweets Sentiment Classifier is a sentiment polarity classifier that classifies each tweet based on the sentiment polarity which it categorizes into three classes: positive, negative or neutral. The relevance tweets classification is performed first; then the tweets sentiment classifier is implemented based on the resulting file from the first classifier. The result of the second classifier is printed in a new column next to the initiative relevance column which is the result of the first classifier. Both Relevance Tweets Classifier and Tweets Sentiment Classifier are grouped under the Text Classification Phase and follow the same steps in building the classification model. Each model covers the following sequential steps:

1) TEXT PREPROCESSING PHASE

This phase is highly important in developing any classification model, since most tweets are unstructured, noisy, and inconsistent. ML algorithms cannot process text directly. Instead, preprocessing steps clean up the text to be machine readable and processable. This preparation phase increases classification accuracy. The preprocessing phase mainly includes four steps: text cleaning, normalization, stemming and stop-word removal.

- Text cleaning. This includes removing each URL link, mentions, numbers, non-Arabic letters, emoticons, multi-spaces, special characters and punctuation marks [20], [21].
- Normalizing the text. This is the process of changing the term or letter to a consistent form. In Arabic languages, the same words and letters can occur in diverse forms, such as with dots and characters appended (i.e. diacritics) to a letter that may or may not be written. For example, replacing the same letters that occur in different forms by one of them, or removing non letters (e.g., Arabic diacritics). Table 1 shows most Arabic normalization conditions [20]–[22].
- Stemming is the process of using algorithms to reduce the word to the possible root form without affecting the meaning, as some words share the same roots and differ only in the affixes. The main aim of stemming is to decrease the number of similar words in the vocabulary, which improves the classification results [21], [22].
- Stop-word removal. This is the process of removing words that do not help in determining the polarity and features using a pre-defined stop-word list as demonstrated in Table 2. It is a critical step and data dependent. Removing words which are important for the classification task will reduce the resulting performance

TABLE 1. Examples of normalizations in Arabic.

Arabic Normalization	Example
Remove Arabic Diacritics	ً, ٰ, ٌ, ٍ, ْ
Remove Tatweel "ـــــ"	Replace the word كريـــم with كريم
Replace Hamza "ء"	Replace ئ, ؤ, ء with ء
Replace Heh "ة"	Replace ة, ه with ه
Replace Alef "ا"	Replace أ, إ, آ with ا

TABLE 2. Examples of Arabic stop words.

Arabic	English
كل	All
هو	He
هي	She
في	In
على	On
من	From

(e.g., removing the negation word will be risky in sentiment classification) [20], 22].

2) FEATURE EXTRACTION PHASE

This phase is concerned with feature extraction and representation. The ''Bag of n-grams'' feature representation strategy is followed. This strategy encompasses two main processes: tokenization and counting. The tokenization process is as follows: each tweet is tokenized into words or tokens using word delimiters such as space or punctuation marks, and each possible token is given an integer id. The tokenized words then form the words' n-gram (wither unigram, bigram or trigram) representation. However, every token (i.e., feature) is represented in the document with an occurrence weight that is usually computed using TF-IDF, regardless of its position in the document. Term Frequency-Inverse Document Frequency (TF-IDF) term weighting, as shown in (1), is used to re-weight or balance among most weighted and less weighted features by the count into float values that are suitable for use by the classifier. Whereas *tf(t,d)* means term-frequency (the number of times that term t occurs in document d) *idf(t)* means inverse document-frequency (the log of the inverse of normalized number of the documents that contain the term t) [22]–[24].

$$\text{tf-idf}\,(t, d) = \text{tf}\,(t, d) \times \text{idf}\,(t) \tag{1}$$

3) CLASSIFICATION PHASE

This phase applies ML algorithms for model training and testing. First, the model is trained based on a various ML algorithm. The one classifier that results in the best performance score is chosen to fit the model using a testing dataset.

C. STATISTICAL ANALYSIS PHASE

This phase is concerned with statistically analyzing the initiative performance based on the classified dataset (i.e., the relevant tweets). As shown in FIGURE 1, the statistical analysis involves the following four phases:

- Tweets Feature Analysis: this statistically measures the overall initiative impact on startups' Twitter accounts based on tweet features.
- Temporal Distribution Analysis: this measures the temporal spread of the initiative across people statistically.
- Sentiment Polarity Classes Ratio: this measures the percentage of sentiment classes in the overall dataset.
- Hashtag Frequencies: this counts each related hashtag frequency over the dataset.

V. EXPERIMENT SETUP

This study makes use of a case study methodology to implement an empirical evaluation of the proposed framework. The startup for this study was the CAREEM transport network company which recently rolled out an initiative aimed at empowering women to work as drivers (captains) across various countries. This initiative was chosen for several reasons. The most significant one was the fast spread of this initiative on Twitter. Empowering women to work with CAREEM was a Twitter trend in Middle Eastern countries, especially in Saudi Arabia. This coincided with the decision to allow women to drive cars in Saudi Arabia. Secondly, this initiative attracted a significant amount of discussion on Twitter between supporters and opposers of this initiative which is perceived as being at odds with the culture in conservative Middle Eastern societies. CAREEM is an app-based transportation platform based in Dubai that was founded in 2012. It operates in more than 15 countries in the Middle East, South Asia, and Africa [25].

This case study contributes to addressing the gap in previous research on. text mining and machine learning fields by focusing on a startup company from an Arab country and on the Arabic language.

A. HARDWARE AND SOFTWARE

The experiments were implemented using Python 3.6 and conducted on a computer with a 2 GHz Intel Core i7 processor and 4 GB of RAM. In addition, the classification models were built using Scikit-learn library.

B. DATASET

The tweets were first crawled using Twitter streaming API through fetching all the tweets that were tweeted and retweeted by the following CAREEM twitter accounts: @Careem, @CareemCare, @CareemKSA, @CareemEGY, @CareemPAK, @CareemUAE, @CareemLEB, @CareemKWT, @CareemBAH, @CareemMAR, @CareemJOR, @CareemQAT, @CareemAUH, @MudassirSheikha and @Abdulla_Elyas.

The dataset crawled using Twitter API has some drawbacks and restrictions. It is limited to accessing only tweets that have been written in the past seven days. Getting older tweets represents an issue. In this experiment, the crawling identified by the initiative date (i.e., from the launch of the initiative until the crawling date), and the reply tweets in this study were an essential part of the dataset since they represented user discussions on the initiative. Therefore, overcoming Twitter API limitations was required. The previous limitations can be overcome using tools which use Twitter advanced search interface to get tweets [26]. The query parameters to such a tool involves a list of keywords, hashtags, exact phrases, Twitter accounts, the start and end date, amongst others. The initiative related keywords and hashtags were listed by exploring and reading the previously crawled dataset using Twitter API and exploring Careem accounts on Twitter.

As a result, the result of crawled tweets was 3,074 from May 1, 2017 to June 1, 2018, including tweets from Careem accounts, the replying tweets for those accounts and the tweets that mention those accounts. Additionally, the tweets that contained keywords, hashtags and exact phrases relevant to the initiative were also crawled. 1000 tweets were labeled manually by two graduate students as a training dataset. Each tweet was assigned two labels: one for the initiative relevance classifier (i.e., Relevant or Not), and one for the sentiment polarity classifier (i.e., Positive, Negative or Neutral). To assess the agreement between two annotators, Cohen's Kappa statistic is used as a reliability measure. The kappa value for sentiment labels is 0.966, which indicates that the degree of agreement between the two annotators is almost perfect. Similarly, the kappa value between initiative relevance is 0.955, which indicates a strong agreement as well. Table 3 shows the class distribution of the training dataset.

TABLE 3. The class distribution of training dataset.

Class	Tweets
Relevancy	
Others	214
Relevant	793
Sentiment	
Positive	323
Negative	471
Neutral	213

VI. EXPERIMENT AND RESULTS ANALYSIS

This section discusses the experiments and results.

A. EVALUATING LEARNING ALGORITHMS

In the first stage of the experiment, which began by training the ML classification models based on the labeled dataset of 1000 tweets, the evaluation process presented the following perspectives:

1) EVALUATION METRICS PERSPECTIVE

For both sentiment classification and subject classification models, the 5-fold cross-validation of the training sets was used for evaluating the classifier. The main job of k-fold cross validation is to ensure that each sample of the data can appear in the test set one time and in the training set (k-1) times to train the classifier model. The most commonly used k-values are 5 and 10. k=5 has been used in this experiment due to the small size of the training dataset. This evaluation was based on four performance measures: Accuracy, Precision, Recall and F1 measure. These performance measures are calculated through a confusion matrix. The Confusion Matrix was used for evaluating the correctness and accuracy of classification problems where the output can be two or more classes. All the performance measures are based on the numbers inside the confusion matrix with these numbers representing the values of the four matrix elements, which are defined as:

- True Positive (TP), the value when the actual class is (True) with the predicted class also (True)
- True Negative (TN), the value when the actual class is (False) with the predicted class also (False)
- False positive (FP), the value when the actual class is (False) and the predicted class is (True)
- False Negative (FN), the value when the actual class is (True) and the predicted class is (False)

2) CLASSIFIERS EVALUATION PERSPECTIVE

This comprised evaluation in terms of comparing various types of classifiers. The main aim was to develop the best model possible by performing a comparative analysis based on the classifier performance. The classifier algorithms that were selected for model training are: Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB), Complement Naïve Bayes (CNB), KNN, Decision Trees (DT), Logistic Regression (LR), Random Forest (RF) and Neural Net. Both SVM and the Naïve Bayes family were the most commonly used algorithms for classification problems in literature. Some works tested KNN in polarity classification while other works used RF in a binary classification problem. Also, DT, LR and Neural Net were tested in this experiment.

3) EVALUATION IN TERMS OF THE PREPROCESSING STEPS PERSPECTIVE

This kind of evaluation is particular to the classification models of Arabic tweets. Arabic text classification is a challenging process, since Arabic language has very complex morphology. The language is made up of 28 letters, with sentences written from right to left, and the form of letters changes according to their position in the word. The following points clarify some of the challenges associated with Arabic text classification tasks [20], [21], [23], [27]:

- The challenge in tokenization step lies in that; Arabic language does not support letter capitalization, with one Arabic word potentially encompassing four tokens and lacking adherence to strict punctuation rules.
- The challenge in Stemming step lies in that it is hard to differentiate between root letters and affix letters in Arabic language and some words have four or five-letter roots. Therefore, stemming is not precise for Arabic.
- The use of negation words which changes the verb meaning to the opposite.
- The Arabic language comes in three forms: the classical Arabic form, modern standard Arabic and colloquial Arabic. Further, Arabic dialects differ from one Arab country to another.
- Most Arabic traded in SM has a considerable percentage of spelling mistakes and repeated letters in some words used to express feelings. For example, the word 'جداااااا' which means something is 'too much'.

To overcome such challenges and improve the classification performance, some previous research [20]–[24] examines the effect of various preprocessing methods. One

study [22] trained both SVM and NB classifiers in four phases, such as without any data preprocess (i.e., raw tweets), using normalization, applying stemmer and removing the stop words. Each phase was run separately based on examining a different combination of word n-grams feature representation. This study [22] also found that preprocessing phases enhance the performance of the NB classifier while both stemming and stop word remover does not improve SVM performance. In a similar manner, other studies such as [23] and [24] sought to investigate the effect of some combination of text representation and preprocessing in Arabic sentiment analysis. The first investigation [23] trained three classifiers in seven representation levels of preprocessing that lie in raw data, stemming, feature correlation and n-gram representation and with some combination of them. The results show that a combination of all preprocessing techniques with feature selection improves the classification accuracy. The second study mentioned earlier [24] also trained three classifiers across three phases of evaluation. It first evaluated weighting schemes (e.g., IDF, TF-IDF, TF). Secondly it assessed the different word n-grams combinations with the best result from the first evaluation. The third phase comprised an evaluation based on various combinations of preprocessing steps with feature selection (i.e., IG) in addition to the best of the previous two evaluation steps. The results indicate that the combination of unigram and bigram give higher accuracy over the different dataset and that the text cleaning and normalization tend to be less effective in accuracy improvements, while stemming techniques and information gain feature selection improves the efficiency significantly. In addition, a study [20] trained three ML classifiers for sentiment analysis with three preprocessing cases as one applied stemming, one applied stop-words remover, and one case combined each of them. While stemming and stemming along with removing stop-words achieved the same results, removing stop words alone yielded the worst results. Similarly, in another study [21], the evaluation was based on combinations of normalization and stemming. The results show that normalization achieved the highest classification accuracy while stemming had lower accuracy.

This experiment followed the approach of the above studies but with extra combinations and techniques described in the following sections. In this study experiment, eight combinations of preprocessing methods were examined for seven supervised ML classifiers; the impact of these methods on the classifier's performance was measured. The applied preprocessing methods are presented in eight levels of combination as in Table 4.

B. CHOOSING THE LEARNING ALGORITHMS

For choosing the best classifier algorithm based on the evaluation perspectives or phases that are discussed earlier, this section presents a detailed analysis of the model training results.

In this experiment, each classifier is built on a pipeline of the following three steps:

TABLE 4. Text preprocessing levels.

Level	Combination of Preprocessing Methods
Level 1	Text Cleaning
Level 2	Text Cleaning + Text Normalization
Level 3	Text Cleaning + Stop Word Removal
Level 4	Text Cleaning + Normalization + Stop Word Removal
Level 5	Text Cleaning + ISRI Stemmer
Level 6	Text Cleaning + Tashaphyne Stemmer
Level 7	Text Cleaning + Text Normalization + ISRI Stemmer
Level 8	Text Cleaning + Text Normalization + Tashaphyne Stemmer

- Feature representation using n-gram. A combination of both unigram and bigram word representations were used to overcome the problem of negation words that hugely affected sentiment classification results 23], [27]. Also, the unigram and bigram combination increased the classification performance in previous research findings [24], [28], while in other research [22] unigram outperformed both bigram and trigram with only 1%.
- Reduce feature dimensions through applying feature selection techniques, using Chi-square and Mutual Information (MI). Results of feature selection are compared with original features.

So, for each of the seven training classifiers, every single run of the preprocessing level has been displayed with three results as shown in Table 5.

TABLE 5. Results (Accuracy, Recall, Precision, F1-Measure, and Time in seconds) of a single model pipeline implementation.

Pre-processing Level	Acc	Rec	Prec	F1	T(S)
Tf-idf + n-gram representation + cleaning	0.80	0.83	0.82	0.82	45
Tf-idf + n-gram representation + cleaning + MI	0.77	0.83	0.82	0.82	45
Tf-idf + n-gram representation + cleaning + Chi-square	0.79	0.83	0.82	0.82	45

Table 5 presents an example of a classification model result with and without the application of feature selection techniques. Feature selection seems to produce the same result for recall, precision, and f1-measure and it affects only the accuracy result. Notably, for most trained classifiers, using feature selection does not improve the accuracy; so, it is excluded from the table of model training result. Further,

Table 6 and Table 7 present only the higher performance measures for every classifier while excluding the other results.

In the following, we discuss the results for each model separately. Notably, the F1 measure was chosen to evaluate and select the classification algorithm due to the unbalanced classes as shown in Table 3.

1) THE INITIATIVE RELEVANCE CLASSIFICATION MODEL

Table 6 shows the training results of the relevance (binary) classification model, with the best result of all performance measures for each classifier highlighted in bold.

TABLE 6. Relevance classification results.

Model	Pre-Processing Level	5-fold cross validation results				T (s)
		Acc	Rec	Prec	F1	
SVM	Level 1	0.80	0.83	0.82	0.82	45
	Level 5	0.79	0.82	0.82	**0.82**	41
	Level 6	**0.82**	0.75	0.73	0.74	30
	Level 7	0.79	**0.84**	**0.83**	**0.83**	39
MNB	Level 5	0.76	**0.83**	**0.70**	**0.76**	31
	Level 8	**0.78**	0.79	0.62	**0.69**	23
CNB	Level 2	0.80	**0.85**	**0.84**	**0.84**	37
	Level 4	0.81	0.83	0.81	0.82	32
	Level 8	**0.82**	0.78	0.79	0.78	24
KNN	Level 3	0.77	**0.80**	**0.76**	**0.78**	30
	Level 3 + Chi-square	**0.79**	**0.80**	**0.76**	**0.78**	30
Decision Tree	Level 2	0.73	**0.82**	**0.79**	**0.80**	29
	Level 6	**0.76**	**0.82**	**0.79**	**0.80**	23
Neural Network	Level 5	**0.82**	0.76	0.73	0.74	37
	Level 6	0.79	**0.82**	**0.83**	**0.82**	28
Random Forest	Level 2	0.78	**0.83**	**0.84**	**0.83**	37
	Level 7	**0.81**	0.77	0.76	0.76	28

TABLE 7. Sentiment classification results.

Model	Pre-processing Level	5-fold cross validation results				T (s)
		Acc	Rec	Prec	F1	
SVM	Level 5	**0.68**	0.62	0.62	0.62	52
	Level 7	0.63	**0.66**	**0.64**	**0.65**	66
MNB	Level 3	**0.58**	0.51	0.56	0.53	26
	Level 4	0.56	**0.59**	**0.68**	**0.63**	25
CNB	Level 4	**0.67**	0.68	0.68	0.68	27
	Level 7	0.66	**0.71**	**0.71**	**0.71**	27
KNN	Level 4 + MI	0.53	**0.51**	0.50	0.50	37
	Level 7 + Chi-square	**0.56**	0.47	**0.56**	**0.51**	28
Decision Tree	Level 1 + Chi-square	**0.54**	0.52	0.51	0.51	29
	Level 5 + Chi-square	**0.54**	0.49	0.51	0.50	23
	Level 6 + Chi-square	0.47	**0.56**	**0.55**	**0.55**	20
Neural Network	Level 2	0.62	**0.68**	**0.70**	**0.69**	41
	Level 8	**0.65**	0.62	0.62	0.62	29
Random Forest	Level 5	**0.64**	0.62	0.61	0.61	26
	Level 7	0.63	**0.66**	**0.66**	**0.66**	27

The overall best result of the F1 measure for the classification model is highlighted in bold with an underline, as it was adopted for evaluating the performance. It can be concluded that CNB classifier outperforms all other classifiers with a 0.84 F1 score. The best results of F1 are achieved when text cleaning is used with text normalization, as both reduce the text noises. Therefore, the CNB classifier is chosen to fit the relevance binary classification as the F1 measure adopted in the model evaluation. However; higher Accuracy results were achieved by the following three classifiers: SVM, CNB, and Neural Net. Additionally, information selection techniques, stop word remover and stemmer techniques seem not to have improved the classification performance in this classification task. It is worth mentioning that all classifier performance results are not conclusive as the performance varies with the dataset and the classification tasks.

2) THE SENTIMENT CLASSIFICATION MODEL

Table 7 shows the case of best performance results for each of the evaluated classifiers highlighted in bold. As with the previous binary classification result, the highest F1 measure is 0.71 when the CNB classifier is used with the same combination along with stemming. The best combination of preprocessing level is text cleaning + normalization + ISRI stemmer. It seems that stemming improves the sentiment classification in this experiment. This case of CNB is chosen to fit the sentiment classification model. It is worth highlighting that all performance results of this sentiment classification are not conclusive, since the performance varies with the study dataset and on the type of classification task. By taking accuracy into account, SVM achieves the highest score at 0.68 when using text cleaning with the ISRI stemmer, followed by CNB with 0.67 in four cases of preprocessing steps, and followed by NN with 0.65 when text cleaning and normalization are used with the Tashaphyne stemmer. In general, information selection techniques reduce the accuracy of the classifiers except for the two classifiers KNN and DT.

Based on previous results of both binary classification and sentiment classification, CNB has been shown to outperform other classifiers. As an interpretation of this performance.

Naïve Bayes (NB) classifier is a probabilistic classifier based on applying Bayes' Theorem with a strong feature independence assumption. These independence assumptions of features assume that the features order is irrelevant; thus the present or absence of one feature concerning a class does not affect any other features in classification tasks [29]–[31]. NB has been widely used in text classification applications; both studies [22], [32] used NB classifiers in sentiment analysis due to its simplicity, effective performance in text classification and great success in sentiment analysis. As well, another study [24] justified the use of NB family classifiers (i.e. Complement NB and Multinomial NB) as these are both widely used in polarity classification and have proven effectiveness in outperforming other classifiers over various datasets. Similarly, another study [33] used NB classifier to classify the sentiment polarity for Arabic and English languages as it is considered the most effective performer in data mining applications. According to other research [30], [34], evidence in literature suggests the effective performance of NB classifiers in classification tasks in general and in sentiment classification in particular. However, Multinomial Naïve Bayes (MNB) and complement Naïve Bayes (CNB) are probabilistic models for NB classification in which MNB is a unigram language model that captures the information of words counts in a document and it performs well with a large vocabulary size. But MNB faces one systemic problem which is that it selects poor weights for decision boundary when one class has more training documents than the others. Consequently, to balance the number of training documents

used per class estimate and to deal with biased training data, a complement version of MNB is proposed that is known as complement Naïve Bayes (CNB) [31], [35]. So, the training results of the present study have proved that the CNB classifier outperforms other classifiers due to its ability to deal with unbalanced classes as shown in Table 3. Moreover, NB classifiers can be trained efficiently through a relatively small training dataset to estimate the features required for classification [13], [30], [36]. Thus, the study's small training dataset (i.e.,1000 tweets) was enough to train the classifier efficiently, whereas in contrast, all other classifiers require large training datasets to estimate the features needed for classification.

3) CLASSIFICATION RESULTS

The overall class distribution results in the dataset (i.e., 3076 tweets) are shown in Table 8.

TABLE 8. The class distribution of both training and testing dataset.

Class	Tweets
Others	789
Relevant	2287
Negative	1817
Positive	827
Neutral	432

C. STATISTICAL ANALYSIS

1) SENTIMENT POLARITY CLASSES RATIO

FIGURE 2 statistically shows the percentage of each sentiment class based on the result of the relevance initiative tweets. As is shown, the negative tweets have the highest percentage with 52.03%, while the positive tweets represent more than the dataset quarter with 33.36% which is a good ratio. The neutral represents the lowest ratio with 14.60% distribution.

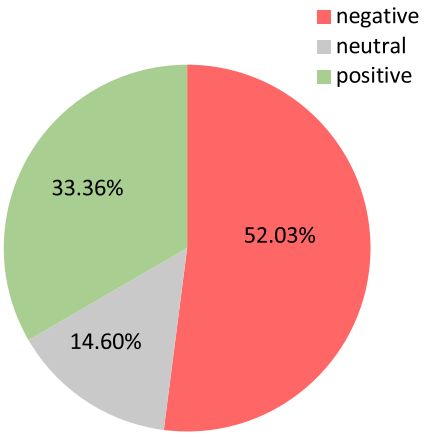

FIGURE 2. Sentiment polarity class distribution.

In Table 9, a random sample of the classified tweet is shown below and the reasons behind the differential percentage of sentiment polarity are inferred. Since the negative tweets represent almost half of the dataset, first five tweets in the table show negative sentiments. All the samples of negative tweets prove that the public opinion was against the idea of empowering women to work as captains in such a transportation company. In tweets 1 and 3, people express the reason for their objection as that the nature of work does not align with women's nature and that such work in public transport will expose women to problems and risks. Another objection noted in tweets with serial No. 2 and 5 is that Careem is exploiting women in the promotion to attract customers and money. In addition, in tweet 4 it is argued that in most world countries women do not work in transportation. In addition, in tweet 4 it is argued that in most of world countries, women do not work in transportation. In the positive tweets, people liked and supported the initiative by some compliment or joining request for such a job. Such examples include tweets 6, 7 and 8. Tweet 8 argued that it is good for women to ride with a woman like her in transport In contrast, the neutral tweets in the sample comprise either inquiries or suggestions regarding the initiative as shown in tweets 9 and 10.

TABLE 9. Tweets sample with polarity of sentiment.

Sr	Tweet in Arabic	English Translation	Polarity
1	أخطر شيء بالحياة إنك تركب المرأة أو الفتاة مع شخص لايعرفها أنا ضد كريم	The most dangerous thing is that, women driving with anonymous person. I'm against Careem.	negative
2	الشركة مستغله والموظفين مستغلين من اسوء الأمثلة لاستخدام المرأة على مر العصور	Its exploiter company and the employees are exploited. It's one of the worst examples of the exploitation of women ever.	negative
3	القرار ذا لازم يلغى ويستبدل بشي يفيد المرآه وظائف شركات نسائية بدل القياده الي بتجلب المشاكل للبلد	This decision must be canceled and replace it by jobs that fit women in the company.	negative
4	عالميا المرأة لا تعمل في النقل في اغلب دول العالم	Universally, women don't work in transportation.	negative
5	بصراحة كريم ليست مهنة كريمة او مناسبة للفتاة.. أنتم هنا تريدون استغلالها فقط لجلب مزيد من الزبائن	To be honest, it's not a decent job or suitable for girl. You want to use them in attract more customers.	negative
6	بادرة كريمة منكم. جزاكم الله عنها خيرا	A decent initiative of you. May you good.	positive
7	سجلوني اول كابتنه	Register me as the first women captain.	positive
8	جميل بناتنا يطلبون كابتنه أفضل. شكرا لكم	It's nice that girls ask other women to ride with them. Thank you.	positive
9	هل أنتم موجودين بدبي وهل عندكم خدمة سيدات في كريم هناك ؟؟	Do you exist in Dubai? Do you have women captain there?	neutral
10	الله يسعدكم جاوبو على سؤالي بعد توظيف النساء عندكم في كريم هل بتكون فيه فنة تسمح لي بطلب كابتنه بدلاً من كابتن؟ @CareemKSA	Answer me please, are we allowed to choose between a man and women captains?	neutral

2) TWEETS FEATURES ANALYSIS

This section examines people's responsiveness in terms of tweet features or attributes based on the initiative relevant

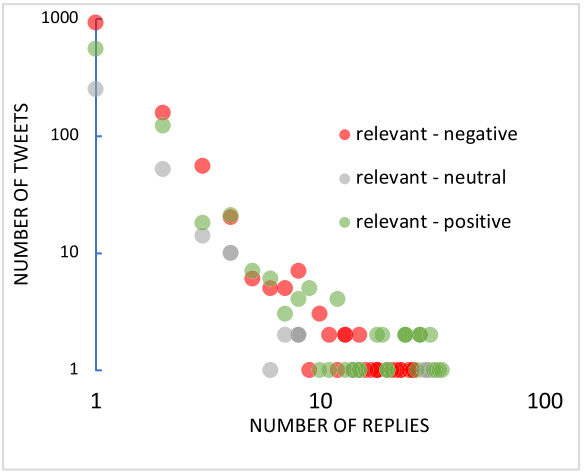

FIGURE 3. Tweet reply frequency.

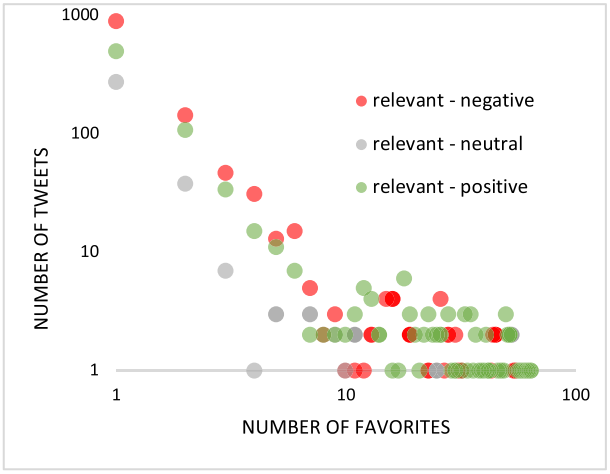

FIGURE 4. Tweet favorite frequency.

TABLE 10. Top five positive tweets having highest number of replies by the user CareemKSA.

Tweet in Arabic	English Translation	Replies
نرحب بأول كابتنه سعودية ما تشتغل في قسم التسويق #المرأة_السعودية_تقود_السيارة	We are welcoming the first Saudi captain, which doesn't work in marketing department. #Saudi_Women_drive_car	796
نرحب بأول كابتنه سعودية ما تشتغل في قسم التسويق #المرأة_السعودية_تقود_السيارة	We are welcoming the first Saudi captain, which doesn't work in marketing department. #Saudi_Women_drive_car	788
نبارك لسيدات الوطن، كابتنه كابتنه. #الملك_ينتصر_لقيادة_المرأة	Congratulations ladies of this country for been captains. #The_King_Empowers_Women_To_Drive	435
نبارك لسيدات الوطن، كابتنه كابتنه. #الملك_ينتصر_لقيادة_المرأة	Congratulations ladies of this country for been captains. #The_King_Empowers_Women_To_Drive	434
أم الحمااااس! بدأ العد التنازلي! باقي ٣٦ أسبوع وتصير فئة سيارة " كابتنة " متاحة لكِ ولعائلتِك	So Exciting! The countdown is started! 36 weeks are remaining for the category of "Captain", it will be available to you and your family.	278

TABLE 11. Top five positive tweets having highest number of favorites by the user CareemKSA.

Tweets in Arabic	English Translation	Fav
نرحب بأول كابتنه سعودية ما تشتغل في قسم التسويق #المرأة_السعودية_تقود_السيارة	We are welcoming the first Saudi captain, which doesn't work in marketing department. #Saudi_Women_drive_car	1556
نرحب بأول كابتنه سعودية ما تشتغل في قسم التسويق #المرأة_السعودية_تقود_السيارة	We are welcoming the first Saudi captain, which doesn't work in marketing department. #Saudi_Women_drive_car	1458
أم الحمااااس! بدأ العد التنازلي! باقي ٣٦ أسبوع وتصير فئة سيارة " كابتنة " متاحة لكِ ولعائلتِك	So Exciting! The countdown is started! 36 weeks are remaining for the category of "Captain", it will be available to you and your family.	797
أم الحمااااس! بدأ العد التنازلي! باقي ٣٦ أسبوع وتصير فئة سيارة " كابتنة " متاحة لكِ ولعائلتِك	So Exciting! The countdown is started! 36 weeks are remaining for the category of "Captain", it will be available to you and your family.	794
نبارك لسيدات الوطن، كابتنه كابتنه. #الملك_ينتصر_لقيادة_المرأة	Congratulations ladies of this country for been captains. #The_King_Empowers_Women_To_Drive	576

tweets. The analysis was done by aggregating the classification results. Aggregation was performed on key engagement features of tweets (reply, favorite and retweet).

As shown in FIGURE 3, the most common response in terms of tweet reply frequency is of positive tweets. On the contrary, the least typical response is higher in negative tweets than in positive and neutral tweets. The top five tweets with a higher number of replies are shown in Table 10. It is noticeable that all top five tweets are from the company official account in Saudi Arabia (i.e., @CareemKSA). These tweets about the initiative which were made by the company for the first time coincided with the decision of allowing Saudi women to drive. However, this higher number of replies proves the higher responsiveness of public in terms of comments or conversation with the official account of Careem in Saudi Arabia.

As shown in FIGURE 4, the most common response in terms of favorite tweet frequency is of positive tweets in which the highest number of tweet favorites relate to positive tweets. On the contrary, the least common response of tweet favorites is higher in negative tweets than in positive and neutral tweets.

Based on Table 11, all the top five tweets of higher favorite number are from the company official account in Saudi Arabia (i.e., @CareemKSA). Accordingly, the higher number of public likes over the study dataset prove the acceptance of and the positive sentiment of public towards such an initiative.

FIGURE 5 shows a similar result to Figure 3 and Figure 4. In general, all tweet key engagement features (i.e., number of Replies, number of Favorites, number of Retweets) are higher in positive tweets, despite the small number of positive tweets

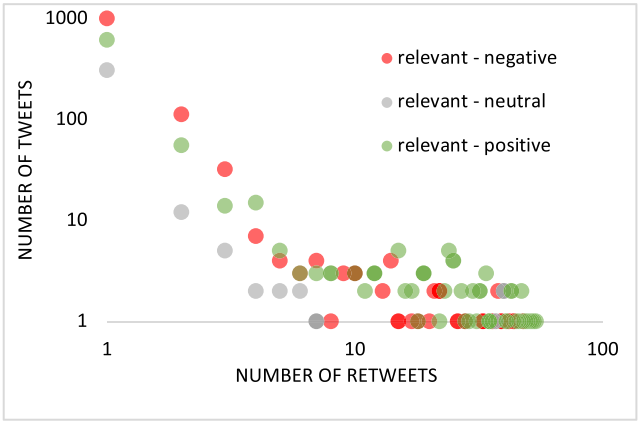

FIGURE 5. Retweet frequency.

(i.e., 763 tweets) compared to negative tweets (i.e., 1190). However, people showed greater response to the positive tweets, while most of the negative tweets drew zero response.

TABLE 12. Top five positive tweets having highest number of favorites by the user CareemKSA.

Tweet in Arabic	English Translation	Retweets
نرحب بأول كابتنه سعودية ما تشتغل في قسم التسويق #المرأة_السعودية_تقود_السيارة	We are welcoming the first Saudi captain, which doesn't work in marketing department. #Saudi_ Women_ drive_car	3450
نرحب بأول كابتنه سعودية ما تشتغل في قسم التسويق #المرأة_تقود_السيارة_السعودية	We are welcoming the first Saudi captain, which doesn't work in marketing department. #Saudi_ Women_ drive_car	3430
نبارك لسيدات الوطن، كابتنه_لقيادة_كابتنه. #الملك_ينتصر المرأة	Congratulations ladies of this country for been captains. #The_King_ Empowers_ Women_ To_Drive	805
نبارك لسيدات الوطن، كابتنه_لقيادة_كابتنه. #الملك_ينتصر المرأة	Congratulations ladies of this country for been captains. #The_King_ Empowers_ Women_ To_Drive	798
أم الحمااااس! بدأ العد التنازلي! باقي ٣٦ أسبوع وتصير فئة سيارة " كابتنه " متاحة لكِ ولعائلتِك	So Exciting! The countdown is started! 36 weeks are remaining for the category of "Captain", it will be available to you and your family.	636

Table 12 shows the top five Tweets in term of the highest retweet number. As with replies and favorites, all the five tweets are from the tweets made by Careem official account (i.e., @CareemKSA). This highest level of retweet indicates the wide spread of the Careem initiative, as retweeting feature proves the content value of the tweet and provides a powerful tool in information dissemination [9]. In addition, the highest number of retweets demonstrate an active user response.

All features like replies, favorite and retweet can be considered as a measure of public active engagement and the

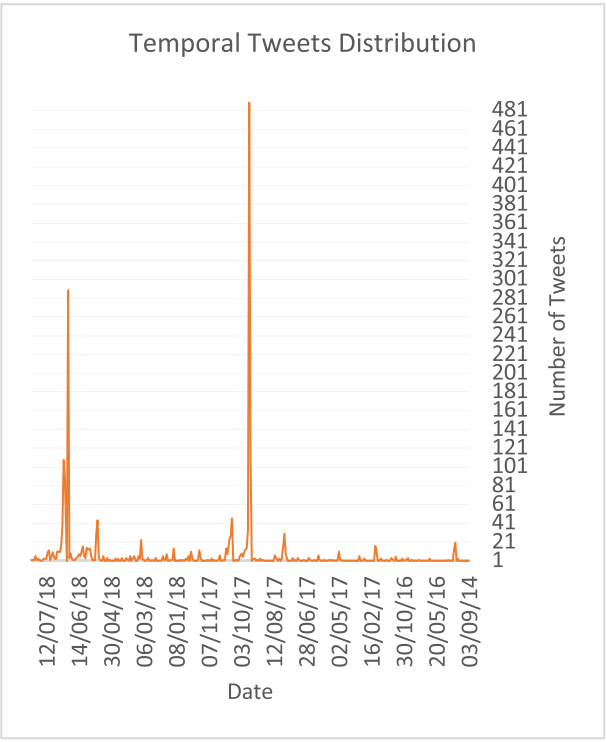

FIGURE 6. Temporal distribution of tweets.

effectiveness of e-WOM [37], [38]. According to Table 10, Table 11, and Table 12, the highest number of these three measures belong to the Careem official account tweets which prove an active customer engagement. Likewise, the highest number of favorites show a good level of customer satisfaction in regard to such an initiative, which can be considered as a measure of public sentiment. Consequently, these tweet engagement features prove that it could be inversely related to the ratio of negative and positive tweets. In addition, it demonstrates that there should not be exclusive reliance on the sentiment tweet ratio to examine public opinion; rather tweet engagement features are also important.

3) TEMPORAL DISTRIBUTION ANALYSIS

FIGURE 6 shows the temporal distribution of respondents. The Temporal analysis was based on the date column in the dataset by creating a pivot table with date order from the oldest to the newest date in the raw and counting the initiative relevant tweet in a column. In tweet crawling, the date of tweets was specified from 1/5/2017, but there are some tweets crawled from 30/9/2014 which contain the initiative keywords. It seems that women started working with Careem in some countries before the launch of the initiative by Careem via Twitter. As in FIGURE 6, the highest distribution of the initiative tweets was from July of 2017 to July of 2018, the period during which the Empower Women initiative was launched through Twitter and disseminated as an initiative within all countries wherein Careem operates. Additionally, FIGURE 6 shows the highest spread distribution of tweets

TABLE 13. Hashtag frequencies of the initiative.

Hashtag	Translation in English	Freq
#الملك_ينتصر_لقيادة_المرأة	#The_King_Empowers_ Women_To_Drive	62
#كريم	#Careem	46
#مبروك_لنساء_الوطن	#Congratulations_to_ homeland_Women	17
#المرأة_السعودية_تقود_السيارة	#Saudi_Women_ drive_car	13
#اوبر	#Uber	13
#استلمي_الكابتنيه	#Receive_Captain	12
#السعودية	#SaudiArabia	12
#قيادة_المرأة	#Women_Driving	6
#السماح_بقيادة_المرأة	#Allow_women_to_drive	5
#حريم_بيتي_لن_يقودوا	#My_women_will_ not_ Drive	5
#كريمة	#Careem (female)	4
#المراه_السعوديه_تسوق	#Saudi_Women_Drive	4
#الشعب_يرفض_قياده_ المراه	#People_refuse_women _driving	4
#فيديو24ثانية	#vidio24second	3
#الرياض	#Riyadh	3
#الدمام	#Dammam	3
#قيادة_المرأة_للسيارة	#Women_Drive_Car	3
#هي_تقود_التغيير	#She_Leads_the_Change	3
#السماح_للمراه_بالقياده	#Empower_Women_ to_Drive_Car	3
#جدة	#Jeddah	3
#السماح_للمرأه_بقيادة_السيارة	#Allow_women_to_ drive_car	2
#كلنا_معهم	#We_are_all_with_them	2
# رويه_2030_تمكين_ المراه	#2030_vision_empower_ women	2
#كريم_يزيد_دخلك	#Careem_Increases_ your_Income	2
#يجب_مقاطعة_كريم_ابوالحريم	#Careen_must_be_boycotted	2
#يلا_نوصلك	#Come_on_we_drive	2
# كابتنة_ كريم	#Careem_ Women_Captains	2
#سواقة_المراة_في_السعودية	#Women_Drive_In_Saudi	2
#يوم_المرأة_العالمي	#International_Women's_Day	2

was between September and October of 2017. This highest distribution coincides with the event of Saudi women being allowed to drive a car which dates to September 2017. Based on that event, Careem welcomed Saudi women to work with and join the company. Accordingly, the discussion on Twitter increased because this kind of work by women is considered incompatible with Saudi culture.

D. HASHTAG FREQUENCIES

Hashtags were extracted from the tweet text of the initiative relevance tweets using the regular expression "\#(w+)". The regular expression extracts all unique hashtags from the tweets. Table 13 shows the initiative relevance Arabic hashtags of the Arabic tweets with the number of occurrences for each one.

The top five Hashtags were selected to examine this wide frequency. The most prevalent one was #The_King_ Empowers_Women_To_Drive that coincided with the granting of permission to drive to Saudi women in the Saudi Kingdom. It was a trend in Saudi Arabia at 26/9/2017. @CareemKSA account and other media accounts participated in this Hashtag such as @MapNewsAR, @Abdulahalsalem and @Ahmed_aleghfeli. The second frequently used Hashtag was the company name (i.e., #Careem) which was widely used as a label for company relevant tweets. Similarly, #Uber was frequently mentioned along with #Careem as a label of company tweets. Uber is a transportation company that has the same business style of Careem, and in March of 2019, Uber acquired Careem for the price of $3.1 Billion. Additionally, both #Congratulations_to_homeland_Women and #Saudi_Women_drive_car coincided with the event of permission to drive being granted to Saudi women. On the other hand, #Receive_Captain Hashtag was relevant to the initiative of empowering women to work as captains in Careem.

VII. CONCLUSION

Large companies often possess fixed budgets and work forces which allow them to easily keep track of business changes and improvements in order to maintain a business presence. On the other hand, startups usually have limited resources, and this can hinder them from tracking the effectiveness of their business operations. The right utilization of SM (i.e., Twitter) analytics techniques can protect companies from failure and support them in generating meaningful, in-depth and reliable business insights to enhance the effective performance of their business operations. The SIRA framework is proposed specifically to support startup founders and entrepreneurs in facilitating the performance of their initiatives by measuring public responsiveness in terms of Twitter activities, customer satisfaction, and the temporal spread. Broadly, the experiment result analysis confirms the effectiveness of such a framework for delivering valuable insights on the initiative based on the Twitter dataset. The experiment was carried out on an Arabic dataset consisting of 3,074 tweets, which were labeled manually with two labels, namely initiative relevance classification and sentiment polarity classes. The dataset is relatively small, which explains why performance measure values are not very high in the two classification models. For both framework models, the experiments were carried out from two evaluation perspectives based on i) seven supervised ML classifiers and ii) different levels of text preprocessing combinations to enhance Arabic classification performance.

The results showed that in both classification models, the CNB classifier achieved a higher F1 measure. In general, simple text cleaning and normalization seemed to significantly increase the binary classification performance, while the applied information selection techniques appeared to have no effect in improving the classifier performance. On the other hand, the results from the sentiment classification model showed a higher F1 measure when text cleaning and normalization were used with ISRI stemmer. Like CNB, most of the higher F1 values for other classifiers were also associated with the use of text cleaning and normalization with stemmer techniques. Like the binary classification model, using information selection techniques seemed to have no effect on the classifier's accuracy except for two classifiers (KNN and DT) although the increase was not high. Analyzing experiment results confirms the effectiveness of SIRA in delivering valuable insights regarding the initiative based on the Twitter dataset. Indeed, people usually express their

opinions through Twitter regarding various topics and issues, not limited to products or services. Consequently, SIRA may be applicable as a Twitter-based analytics framework in various domains and for diverse purposes.

REFERENCES

[1] W. He, H. Wu, G. Yan, V. Akula, and J. Shen, "A novel social media competitive analytics framework with sentiment benchmarks," *Inf. Manage.*, vol. 52, no. 7, pp. 801–812, Nov. 2015.

[2] R. A. Abbasi, O. Maqbool, M. Mushtaq, N. R. Aljohani, A. Daud, J. S. Alowibdi, and B. Shahzad, "Saving lives using social media: Analysis of the role of Twitter for personal blood donation requests and dissemination," *Telematics Inform.*, vol. 35, no. 4, pp. 892–912, Jul. 2018.

[3] M. A. Jarwar, R. A. Abbasi, M. Mushtaq, O. Maqbool, N. R. Aljohani, A. Daud, J. S. Alowibdi, J. Cano, S. García, and I. Chong, "CommuniMents: A framework for detecting community based sentiments for events," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 2, pp. 87–108, Apr. 2017.

[4] S. Stieglitz and L. Dang-Xuan, "Political communication and influence through microblogging-an empirical analysis of sentiment in Twitter messages and retweet behavior," in *Proc. 45th Hawaii Int. Conf. Syst. Sci.*, Jan. 2012, pp. 3500–3509.

[5] T. Rao and S. Srivastava, "Modeling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments," in *Proc. 5th Annu. ACM Web Sci. Conf. (WebSci)*. New York, NY, USA: ACM, 2013, pp. 336–345.

[6] P. Capriotti and L. Ruesja, "How CEOs use Twitter: A comparative analysis of global and Latin American companies," *Int. J. Inf. Manage.*, vol. 39, pp. 242–248, Apr. 2018.

[7] S. Lugovi and W. Ahmed, "An analysis of Twitter usage among startups in Europe," Dept. Inf. Commun. Sci., Fac. Hum. Social Sci., Univ. Zagreb, Zagreb, Croatia, Tech. Rep., 2015, doi: 10.17234/INFUTURE.2015.32.

[8] M. S. Shabbir, M. S. Ghazi, and A. R. Mehmood, "Impact of social media applications on small business entrepreneurs," *Manage. Econ. Res. J.*, vol. 2, p. 1, 2016.

[9] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Amer. Soc. Inf. Sci.*, vol. 60, no. 11, pp. 2169–2188, Nov. 2009.

[10] C. Holsapple, S.-H. Hsiao, and R. Pakath, "Business social media analytics: Definition, benefits, and challenges," in *Proc. 20th Amer. Conf. Inf. Syst.*, Savannah, GA, USA, 2014. [Online]. Available: https://pdfs.semanticscholar.org/d7d7/1ec49476e54a350e9091087345dcd3d7866c.pdf

[11] B. A. Kitchenham, "Guidelines for performing Systematic Literature Reviews in software engineering," EBSE, London, U.K., Tech. Rep. EBSE-2007-01, 2007.

[12] C. Okoli and K. Schabram, "A guide to conducting a systematic literature review of information systems research," Work. Papers Inf. Syst., Sprouts, USA, Tech. Rep., 2010. [Online]. Available: http://sprouts.aisnet.org/10-26

[13] T. Antretter, I. Blohm, and D. Grichnik, "Predicting startup survival from digital traces: Towards a procedure for early stage investors," in *Proc. Int. Conf. Inf. Syst. (ICIS)*, San Francisco, CA, USA, 2018. [Online]. Available: https://pdfs.semanticscholar.org/31dc/753345d5230e421ea817dd7dcdd352e87ea2.pdf

[14] B. Sharchilev, M. Roizner, A. Rumyantsev, D. Ozornin, P. Serdyukov, and M. De Rijke, "Web-based startup success prediction," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: ACM, 2018, pp. 2283–2291.

[15] A. Krishna, A. Agrawal, and A. Choudhary, "Predicting the outcome of startups: Less failure, more success," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 798–805.

[16] J. Kuruzovich and Y. Lu, "Entrepreneurs' activities on social media and venture financing," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 1944–1952.

[17] M. M. Gambardella, "The influence of a start-up process on the entrepreneurs' emotions, deduced by their Twitter accounts," M.S. thesis, Univ. Católica, Lisbon, Portugal, 2017, doi: 10400.14/22036.

[18] Y. Song and G. Zeng, "Uncover successful entrepreneurs' crowdfunding behaviors through Twitter," in *Proc. 5th Int. Conf. Collaborative Innov. Netw. (COINs)*, Tokyo, Japan, 2015. [Online]. Available: http://tokyo15.coinsconference.org/proceedings/COINS15_35_Song_Zeng.pdf

[19] S. S. Alsheikh, K. Shaalan, and F. Meziane, "Consumers' trust and popularity of negative posts in social media: A case study on the integration between B2C and C2C business models," in *Proc. Int. Conf. Behav., Econ., Socio-Cultural Comput. (BESC)*, Oct. 2017, pp. 1–6.

[20] C. Laurell and C. Sandström, "Analysing uber in social media-disruptive technology or institutional disruption?" *Int. J. Innov. Manage.*, vol. 20, no. 5, 2016, Art. no. 1640013.

[21] D. Ulloa, P. Saleiro, R. J. F. Rossetti, and E. R. Silva, "Mining social media for open innovation in transportation systems," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 169–174.

[22] R. Ismail, M. Omer, M. Tabir, N. Mahadi, and I. Amin, "Sentiment analysis for arabic dialect using supervised learning," in *Proc. Int. Conf. Comput., Control, Electr., Electron. Eng. (ICCCEEE)*, Aug. 2018, pp. 1–6.

[23] R. M. Sallam, H. M. Mousa, and M. Hussein, "Improving arabic text categorization using normalization and stemming techniques," *Int. J. Comput. Appl.*, vol. 135, no. 2, pp. 38–43, Feb. 2016.

[24] N. F. B. Hathlian and A. M. Hafez, "Subjective text mining for arabic social media," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 2, pp. 1–13, Apr. 2017.

[25] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text," *J. Inf. Sci.*, vol. 40, no. 4, pp. 501–513, Aug. 2014.

[26] T. Khalil, A. Halaby, M. Hammad, and S. R. El-Beltagy, "Which configuration works best? An experimental study on supervised arabic Twitter sentiment analysis," in *Proc. 1st Int. Conf. Arabic Comput. Linguistics (ACLing)*, Apr. 2015, pp. 86–93.

[27] Careem Website. *Our Story*. Accessed: Dec. 14, 2019. [Online]. Available: https://www.careem.com/en-ae/our-story/

[28] M. Abdullah, M. Almasawa, I. Makki, M. Alsolmi, and S. Mahrous, "Emotions extraction from arabic tweets," *Int. J. Comput. Appl.*, pp. 1–15, Jun. 2018.

[29] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2012, 9.

[30] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[31] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *J. Adv. Inf. Technol.*, vol. 1, no. 1, pp. 4–20, 2010.

[32] S. Bhuta and U. Doshi, "A review of techniques for sentiment analysis of Twitter data," in *Proc. Int. Conf. Issues Challenges Intell. Comput. Techn. (ICICT)*, Feb. 2014, pp. 583–591.

[33] S. M. Vohra and J. B. Teraiya, "A comparative study of sentiment analysis techniques," *J. JIKRCE*, vol. 2, no. 2, pp. 313–317, 2013.

[34] M. E. M. Abo, N. A. K. Shah, V. Balakrishnan, and A. Abdelaziz, "Sentiment analysis algorithms: Evaluation performance of the Arabic and English language," in *Proc. Int. Conf. Comput., Control, Electr., Electron. Eng. (ICCCEEE)*, Aug. 2018, pp. 1–5.

[35] A. Sharma and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," in *Proc. ACM Res. Appl. Comput. Symp. (RACS)*, 2012, pp. 1–7.

[36] L. Jiang, Z. Cai, H. Zhang, and D. Wang, "Naive Bayes text classifiers: A locally weighted learning approach," *J. Experim. Theor. Artif. Intell.*, vol. 25, no. 2, pp. 273–286, Jun. 2013.

[37] L. Zhang, "Sentiment analysis on Twitter with stock price and significant keyword correlation," Texas ScholarWorks, Univ. Texas Austin, Austin, TX, USA, Tech. Rep., 2013. [Online]. Available: http://hdl.handle.net/2152/20057

[38] W. Fan and M. D. Gordon, "The power of social media analytics," *Commun. Acm*, vol. 57, no. 6, pp. 74–81, 2014.

[39] N. Weerawatnodom, N. Watanapa, and B. Watanapa, "Features of marketer-generated content tweets for electronic word of mouth in banking context," *KnE Social Sci.*, vol. 3, no. 1, p. 82, Mar. 2018.