Journal of Computer Science

Original Research Paper

Twitter Sentiment Analysis for Reviewing Tourist Destinations in Saudi Arabia using Apache Spark and Machine Learning Algorithms

[1]Wala Awadh Alasmari and [1,2]Hoda Ahmed Abdelhafez

[1]*Department of Information Technology, College of Computer and Information Sciences, Princess Nourah University, Saudi Arabia*
[2]*Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt*

Article history
Received: 16-12-2021
Revised:05-03-2022
Accepted:15-03-2022

Corresponding Author:
Hoda Ahmed Abdelhafez
Department of Information Technology, College of Computer and Information Sciences, Princess Nourah University, Saudi Arabia
Email: hodaabdelhafez@gmail.com

Abstract: The appearance of big data has created new challenges for data analysis teams especially dealing with unstructured data in text form. Many applications increasingly include a large amount of this type of data. Example of such data is data collected from Twitter. Adequate use of Machine Learning (ML), big data tools and social media platforms can solve several problems. The aim of this research is to apply sentiment analysis using Arabic tweets of tourism in Saudi Arabia and determine the most visited places. Ara Senti corpus was used as the labelled data to perform machine learning for sentiment analysis to deal with the Arabic morphology. The three-classes classification (Positive, Negative, or Neutral) was performed using Decision Tree, Random Forest, Logistic Regression and Naïve Bayes. The results showed that the highest performance achieved was 86% using Logistic Regression with Term Frequency–Inverse Document Frequency (TF-IDF) representation and Naïve Bayes with Bag-of-Words model compared with both random forest and decision tree. The trainable classifier was applied to predict classes on collected data from Twitter for reviewing Kingdom of Saudi Arabia (KSA) destinations to finally present a rating of the most visited places on KSA. There are five most visited places in Saudi Arabia (Riyadh, Alula, Hail, Taif and Tabuk).

Keywords: Twitter, Big Data, Machine Learning, Sentiment Analysis, Tourism

Introduction

Social media is one of the most popular interactive media that breaks down the barriers of society's rules and people start making decisions based on it. As social media and Twitter are specifically growing fast, the amount of their data is growing as well. Analyzing these large data volumes becomes increasing difficult for private and public organizations.

The monthly active Twitter users are 330 million around the world, 40% of them are daily online (Lin, 2020). October 2019, the fifth highest number of Twitter users (10.09 million) is in Saudi Arabia according to Statista reports (Statista, 2021). Moreover, KSA has the largest amount of Internet users who are online on the Twitter. Around 80% of the users have access twitter through smartphones to obtain rich Spatio-temporal information (Omnicore, 2021).

Twitter is a powerful informational tool for broadcasting information and a rich source of opinion texts on various topics: Business, economic, politics, social and tourist. This has enthused the interest in using machine learning and big data in the research community to study this rich linguistic resource. Mubarak and Darwish (2014) collected numerous Arabic tweets from Twitter; the dataset of 175 million Arabic tweets was collected then after filtering tweet user location, a subset of 6.5 million tweets was classified corresponding to the tweet's dialect, the authors found that 61% of the tweets were in Saudi dialect, followed by 13% Egyptian and 11% Kuwaiti. This indicates the huge presence of the Saudi community on Twitter (Mubarak and Darwish, 2014).

In recent years, the growth of Saudi's tourism industry increased significantly, but it has also highlighted some key issues the kingdom must resolve to make it more attractive for international visitors. According to Saudi Commission for Tourism and Antiquities' (SCTA) and Tourism Information Research Centre (MAS), the number of inbound visitors of Saudi Arabia was ranged from 11 to 17 million from year 2006 to year 2011. At the end of the year 2020, the number of tourists raised from 20.8 million to 45.3 million. This means a significant investment required in both private and the government sectors (MAS, 2012). Therefore, the tourism sector in Saudi Arabia has witnessed tremendous growth in visitors and a significant increase in domestic tourism. The Saudi government plan is to increase the number of visitable heritage destinations from 241 to 447 (NTP, 2021).

Twitter allows users to express themselves and share their opinions with others worldwide (Omnicore, 2021). Thus, there is a massive potential of Twitter data analytics, which led many researchers to use this potential to provide business/social outcomes. However, many areas are still unexplored. There is a research gap in exploring Twitter data about tourism destinations in Saudi Arabia using the Arabic language. To overcome the research gap, this study used the combination of big data technologies and the massive data provided by Twitter. It attempted to provide Twitter sentiment analysis of Arabic tweets concerning tourism in Saudi Arabia to support tourism decision makers. The most challenge in the study is dealing with Arabic text with its complex structure of the words and morphology as well as the Saudi dialect. To overcome this challenge labelled data is used as an annotated corpus of Arabic tweets to fit the collected dataset and handling both modern standard Arabic and Saudi dialect.

The Twitter Developer Application Programming Interface (API) consists of many different endpoints, but the most central one in this study is the Search endpoint. Twitter offers three tiers of search APIs: Standard, premium and enterprise. This research used the Premium tier (Full-Archive/Sandbox), which provides access to the historical tweets from 2006 until now. The maximum number of returned tweets was 500 tweets per request. Furthermore, these tweets were filtered by specific parameters in the API call, such as sender, recipient, or posting date (Campan *et al*., 2018).

Machine learning and artificial intelligence have a major added value in the tourism industry, as well as many other fields (Verbraeken *et al*., 2020). In this regard, the promotion of intelligent tourism is highly valued; it can help in developing the tourism industry and improving its services. There are other technologies used, such as big data (Miah *et al*., 2017), cloud (Zhiqiang and Changguo, 2016) and Apache Spark (Ntaliakouras *et al*., 2019).

Alomari *et al*. (2021) defined big data term as a large growing dataset that include heterogeneous formats: Structured, unstructured and semi-structured data. These new formats need advanced and powerful technologies to deal with their heterogeneity and complexity. One of these technologies is Apache Spark. It is an open-source big data processing framework that is designed for speed and facilitate the sophisticated analysis on massive amount of data.

The Natural Language Processing (NLP) enables a machine to handle a natural language and translates it into a machine-readable format (Rajput, 2020). Sentiment analysis is an important field of natural language processing; it is also known as an opinion mining (Omari and Al-Hajj, 2020). Sentiment analysis is the process of analyzing people's feelings, perceptions, behaviors and emotions about things such as the visited places, the obtained products and company services. Many of the sentiment analysis research studies used machine learning and social media data such as Twitter. Alaei *et al*. (2019) suggested that adopting Naïve Bayes analysis could provide fast and accurate probabilistic output regarding the impact of human emotion in the tourism sector.

The rest of this research study is organized as follows: Literature reviews of recent studies that relate to the sentiment analysis field, including background, related works and research gap. The research methodology section includes the outline of natural language processing for Arabic text and different ML algorithms: Decision Tree, Random Forest, multinomial logistic regression and Naïve Bayes (NB). The data and analysis section focus on data collection and pre-processing. Implementation and results section focus on presenting the tools used, the implementation, the evaluation of the models and results. The last section represents the conclusion and future work.

Related Works

Analysis the data generated from Twitter provides unexpected opportunity to enrich the tourism sector. The most known methodology is sentiment analysis, which aims to collect and analyze (using ML) people's opinions. The notable works are review below.

The study using Twitter data in Arabic language was in the field of healthcare. After the COVID-19 pandemic, the studies increased about how the pandemic affects society from social media. Alomari *et al*. (2021) used Twitter Arabic Data and Distributed Machine Learning to identify government pandemic measures against COVID-19 as well as public concerns. The authors developed a software tool that used unsupervised Latent Dirichlet Allocation (LDA) machine learning and natural language processing to analyze Arabic Twitter data to detect government pandemic steps and public concerns related to the COVID-19 pandemic. From 1 February 2020 to 1 June 2020, 14 million tweets were collected from the Kingdom of Saudi Arabia. The result showed the Twitter media's effectiveness in identifying the significant event, government actions and public issues.

Alhajji *et al*. (2020) introduced Tweets' sentiment analysis of governmental preventive steps to contain COVID-19 in Saudi Arabia. The study focused on an Arabic annotated dataset about COID-19, which consisted of 53,127 tweets. The authors collected the data from a particular hashtag in Saudi Arabia about the curfew and the preventive measures, then applied Naïve Bayes machine Learning model and (NLTK) library Python.

Universities utilize social media in educational practice to improve their teaching processes, learn about experiences and analyze opinions. Al-Rubaiee *et al*. (2016) applied sentiment analysis of Arabic Tweets in e-Learning. The study presented an Arabic text classification implementation regarding King Abdul-Aziz University students' opinions using Support Vector Machine (SVM) and NB algorithms. This study collected very small dataset around two thousand tweets by different students from King Abdul-Aziz University students in 2016. The authors implemented several criteria to collect the datasets; These criteria include: (1) Tweets without hashtags, (2) Tweets without links/URLs, (3) eliminating duplicated tweets and (4) Tweets without special characters. Other Tweets stored in a reserved database and marked the tweet as negative, positive, or neutral.

Another study conducted by Alruily and Shahin (2020) focused on sentiment analysis of Twitter data for Saudi Universities. The authors classified tweets to develop a sentiment analysis system for analyzing Tweets generated by Saudi Twitter users. The classification method was a K-Nearest Neighbors classifier (KNN), SVM and NB. The dataset consisted of 600 K tweets collected from comments, after applying the classification model to remove irrelevant words, classified around 60 K tweets as positive and negative.

Duwairi (2015) presented sentiment analysis for dialect words in Jordan's Arabic language. This research focused on studying the comments and the reviews in a Twitter platform to determining whether a tweet was positive, negative, or neutral. The author applied SVM and NB classifiers. The total dataset collected around 22550 tweets using Twitter API search. The results showed that replacing dialectical terms with their Modern Standard Arabic (MSA) equivalents.

Aldayel and Azmi (2016) showed sentiment analysis for Arabic tweets using a hybrid scheme. The aim of this research was to develop a sentiment analysis seeks to identify the tweets' division, whether positive or negative. The authors addressed the social issues in Saudi Arabia in many general topics posted in Twitter. The dataset was around 50 K tweets, which collected using Twitter API Search. The classifier was lexical and SVM.

Other research studies used data from different websites related to tourism sector. The review of the tourist's guests about hotels is an important issue for tourism. Usually, when the visitor checkout from a resort or the hotel, a review about the place post on hotel websites so that other visitors can benefit from this review. On the other side, the place owner can review this feedback to enhance guests' services. One of research studies conducted by Alosaimi *et al*. (2020) on this matter to help hotels assess the administrative and operational staff's quality and to evaluate customers' satisfaction with the provided services. The authors presented an approach based on unsupervised machine learning methods to discriminate between positive and negative reviews. The dataset of 4604 Arabic reviews from 121 Saudi hotels collected from the TripAdvisor website. Methodology steps were collecting the dataset, pre-processing the data, features extraction by TF-IDF, clustering by K-means and Hierarchical algorithms, evaluation and then the results sentiment analysis into positive and negative class for collected Arabic reviews of Saudi hotels.

Chen *et al*. (2020) used sentiment classification to study and analysis online travel review texts. The proposed method used Microsoft Knowledge Graph to extract keywords from online travel review text and generate a concept list of keywords. To create an efficient online sentiment classification model for travel review text, the authors applied keyword extraction, classification labelling and machine learning-based sentiment classification methods. The method of sentiment classification was SVM. The dataset source in this study was from the TripAdvisor website, contained 20 K reviews text datasets. The study result was tourist opinions on travel destinations obtained from online travel review texts.

In summary, many studies demonstrate the massive potential of Twitter data analytics, which led many researchers to use this potential to provide business/social outcomes. However, some areas like tourism industry are still unexplored. There is a rare research studies using sentiment analysis of Arabic tweets about tourism destinations especially in Saudi Arabia. Thus, this research attempted to focus on analysis Arabic tweets concerning tourism in KSA.

Research Methodology

This section discusses methods and techniques for pre-processing textual data including the cleaning and the normalization of text, types of transformations, presenting machine learning algorithms as well as evaluation methods.

Tweet’s classification based on their semantic orientation (Positive, Negative or Neutral) provides descriptive of visitors’ opinions for the most visited places in the KSA. Such analysis is great interest to a variety of stakeholders, including tourism organizations, ministries of tourism, travel companies. This analysis requires the use of machine learning methods. To accomplish the sentiment analysis classification task, the methodology framework is presented in Fig. 1.

Data Preparation

Data preparation is crucial in text analysis and information extraction (Soliman *et al*., 2017). Pre-processing a text is basically putting it in an appropriate form. A series of techniques that could be applied in a general way to make a text useful are text normalization, Noise removal, removing stop-words, stemming, lemmatization, tokenization and features extraction. Text normalization is the process of transforming text into a standard form.

DOI: 10.3844/jcssp.2022.215.226

One of the important characteristics of the Arabic language is the morphology, which plays an important role. Arabic letters have different shapes depending on where these letters are in a word. The Arabic language's complex word structure and morphology have made processing Arabic text a major challenge (Duwairi and El-Orfali, 2014). Examples of possible actions that can be executed for Arabic text normalization are (1) Strip Harakat from Arabic word except Shadda and (2) Reduce the Tashkeel, by deleting evident cases (Zerrouki, 2021).

Noise removal is the process of removing characters, numbers and pieces of text that may interfere with the analysis (Boujou *et al*., 2021). Noise removal is one of the most required steps in text pre-processing. Special characters and numeric characters can be replaced with space. This step is very important because punctuation does not add any additional information or value. Consequently, all these instances will aid in the reduction of data size and increase the calculations efficiency.

Another important step of the text pre-processing is removing the stop-words from the text (Boujou *et al*., 2021). Stop-words appear too frequently in any type of text. This particularity means that their presence does not provide any useful information for the classification of the text. The presence of these words can, on the contrary, produce noise that complicates accurate classification. Therefore, it is preferable to remove these words to improve the classification capacity of the model that will be used later. Some stop-words of the Arabic language from the NLTK python library are: ['إلى''إذن' ,'إذا' ,'إن' ,], which means [To, So, If, That].

Stemming means extracting for each token a root (Duwairi and EL-Orfali, 2014). If there are two words, one conjugated or tuned and the other one represents its original terminology, they cannot be considered similar words since their spellings are different. However, if the root of these two words is similar then consider them as the same meaning. The root of a word is obtained by applying a stemming algorithm. Light stemming removes only prefixes and suffixes to extract the root. An example for return the root of the words. is: ['علمتهن','متعلم','تعلمنا'] [taught them, learner, learned them] are considered as different words. These words can be seen by the computer as the same term 'علم' (learn). Or as two varieties 'متعلم' ,'علم' (learner, learn) using the light stemming (Zerrouki, 2021).

Lemmatisation is quite similar to the stemming. It can be done instead of applying stemming. Lemmatization is the process of assigning a corresponding lemma to each surface form of a word in a text, that is, the canonical form of the word as it appears in a dictionary (Freihat *et al*., 2018). Since of the rich morphology of Arabic, lemmatization is a complex task. Lemmatization has no significant advantage over base acquisition for text search and classification purposes.

Tokenization: Tokenization is the process of splitting the text into words called tokens (Words, numbers, punctuation marks) (Bird *et al*., 2009). Generally, tokens are separated using punctuation and spaces. This task essentially divides the text into the basic structures for further analysis. These structures can be words (monograms) sets of two or more adjacent words (bigrams or mgrams) phrases, symbols, or other basic structure that provides useful information for classification. Spaces and punctuation marks in the original text could, or could not, be included in the resulting list of tokens.

Features Extraction focuses on representing list of tokens as numerical feature vectors which can be fed into machine learning algorithms directly or after further processing. A list of tokens was obtained after processing the initial text. This list is currently incomprehensible for the algorithms that need to receive numerical vector representations of the entities to be classified (MLlib, 2021). The vector representation includes transforming each document into a sequence of numbers, in which each number corresponds to a word of the vocabulary of the set of documents or corpus. Bag of words and TF-IDF (Term frequency–inverse document frequency) are two common feature extraction methods.

Bag of Words: Creating a vocabulary of all unique words from the corpus and then creating a matrix of features by assigning a separate column for each word, while each row corresponding to a document is one of the most basic methods for transforming tokens into a collection of features (text) (MLlib, 2021). Generating a vector of tokens in randomized order is losing the order of occurrence of words, which represents a major disadvantage instead of using individual words (i.e., unigrams); this problem can be solved by considering N-grams mostly bigrams. However, using N-grams will result in a huge feature vector proportional to the vocabulary size, making computations more difficult.

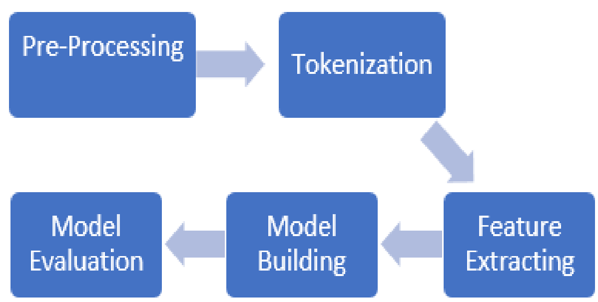

Fig. 1: Methodology framework

Term Frequency-Inverse Document Frequency (TF-IDF): For each word in a text, the main concerns are its frequency of appearance in the text and its frequency of appearance in the corpus. The TF-IDF method is used to reflect the importance of a term to a document in the corpus (MLlib, 2021). It is a method of converting documents into vectors in which the vector represents the importance of a term to a document in the corpus. Terms with a high frequency in the document would have a high TF, but if a term has a high frequency in the corpus, it is necessary to reduce by IDF. The TF-IDF of a term that appears in all documents in the corpus is equal to 0. The more frequently a word appears in a text, the more important it is. However, the more frequently a word appears in the corpus, the less important it is. The inverse document frequency is a numerical value that indicates how much information a word contains

Machine Learning Algorithms

Machine Learning is a sub-field of Artificial Intelligence (AI), which aim to train a set of algorithms on large amounts of data to be able to classify or predict future data (the accuracy depending on the quantity and quality of the data), (Kubat, 2017). This research study applied four types of machine learning algorithms.

Multinomial Logistic Regression

Multinomial logistic regression is a multi-class classification extension of logistic regression. It is a supervised classification algorithm. It is an old statistical classification model that has been rediscovered and has recently gained great popularity due to its good performance in automatic classification. LR is a two-class classification algorithm that uses a binomial probability distribution function to model the target. The negative class or outcome is mapped to 0 and the positive class or outcome is mapped to 1. The fit model predicts the probability of a case belonging to class 1. Multinomial logistic regression is a type of LR that predicts a multinomial probability for more than two classes. The LR is a particular case of the generalized linear model, which is use to predict the probability of the answer 1 rather than the value directly (0 or 1). Since this model is very simple, there is little risk of overfitting and the results tend to have good generalization power. The conditional probabilities of the predicted classes $k \in 1,2,\ldots,K$ are modelled using the soft max function. The weighted negative log-likelihood is reduced using a multinomial response model, with elastic net penalty to restrain overfitting (MLlib, 2021).

Decision Trees

It is called Classification and Regression Trees (CART) (Jo, 2021). The acronym CART corresponds to two distinct situations depending on whether the variable to be explained, modelled, or predicted is discrete (classification) or continuous (regression). CART is nonparametric and unsupervised machine learning algorithm. The advantage of this algorithm is its relatively simple explanatory power since the obtained predictions are presented in an easy-to-interpret graphical form and constitute an effective aid for decision support. These predictions are based on a recursive sequence of division rules. Decision tree is easily interpreted; however, its predictive ability is almost exceeded by other classification models. This characteristic limited the use of decision tree until the early 2000 s, then it was taken up as the basis of a new technique, called the decision forest. This new technique uses a combination of decision trees and statistical theory to reduce the variance of the classifier by calculating the average of a set of decision trees by generating binders with a very good predictive ability.

Random Forest

The random forest model is based on the theory of the CART algorithm explained earlier. It is the natural evolution of CART and, as it often provides better predictions. RF consists of a set of independent decision trees (Breiman, 2001). This model belongs to the family of model aggregations; it is in fact a particular case of bagging (bootstrap aggregating) applied to CART.

The Algorithm is based on two principal processes: Tree bagging and feature sampling. Applying these processes, the algorithm produces several trees, each individual tree provides a prediction. When the desired number of trees has been simulated, the prediction can be obtained in two different ways depending on the approach used (classification or regression). In the same way as for the CART algorithm, the algorithm chooses the most represented class in a classification problem, or it calculates the average of the outputs if it is a regression. The model uses a set of trees to calculate the "average" forecast value and bagging to reduce variance, thus rendering a model that will have a much better generalization capability than individual trees. Since the Random Forest model provide better performance and more robust than a single decision tree. However, using this algorithm may cause the problem of overfitting (that means the model fits the training data very well, but the model fails to generalize for new input data) (Manorathna, 2020).

Naïve Bayes

The naive Bayes is a probabilistic classifier based on Bayes' theorem and the "naive" assumption of independent features. Naive Bayes classifier is simple, robust and widely used. This method is often used in document categorization and classification (Yuliana and Erlangga, 2017). The objective is to estimate the posterior probability of each class among the examples P (label features) and assign to it the most probable class (Bird *et al*., 2009).

DOI: 10.3844/jcssp.2022.215.226

Because of its effectiveness, the naive Bayes algorithm is widely used for text classification tasks. Depending on the words representation to create the input vectors, the used distribution is: (1) Word present: Binary distribution, (2) word occurrence: Multinomial distribution and (3) Frequency (TF-IDF): Gaussian distribution

Evaluation

To explain the model's performance, the following evaluation metrics are used: F1 score, Recall and Precision (MLlib, 2021).

Recall is a measure that indicates how many accurate positive predictions were made from all possible positive predictions (Bird *et al*., 2009).

$$Weighted\ Recall = \frac{1}{N}\sum_{i=0}^{N-1}\delta\,(y_i - c)$$

$$Recall\,(c) = \frac{True\,Positives\,(c)}{True\,Positives\,(c) + Flase\,Negatives\,(c)}$$

Precision is a metric that measures how many accurate positive predictions have been made.

$$Weighted\ Precision = \frac{1}{N}\sum_{c\,in\,c} Precision\,(c) \times \sum_{i=0}^{N-1}\delta\,(y_i - c)$$

$$Precisions\,(c) = \frac{True\,Positives\,(c)}{True\,Positives\,(c) + False\,Positives\,(c)}$$

F1 Score: F-Measure combines precision and recall into a single metric that captures both properties:

$$Weighted\ F1.Score = \frac{1}{N}\sum_{c\,in\,c} F.1score\,(c) \times \sum_{i=0}^{N-1}\delta\,(y_i - c)$$

$$F1score_c = \frac{2 \times Percisions\,(c) \times Recall\,(c)}{Precisions\,(c) + Recall\,(c)}$$

where:

- C the list of classes
- $\delta(x)$ delta function: $\delta(x) = \begin{cases} 1 if\ x = 0 \\ 0 otherwise \end{cases}$

Big Data and Twitter

The use of social media has dramatically increased. Large amounts of data are created every day by millions of users. Such volume of data can be efficiently analyzed using big data tools and Machine Learning algorithms (ML) to explore and extract meaningful information from raw records (Alaoui and Gahi, 2019). The extracted information is extremely important for companies and organizations to take better decisions and manage their businesses. One of the most used of social media platforms is Twitter where users post comments to express their interests, reviews and opinions. Using the combination of Big Data and ML can automatically identify the semantic orientation of a given opinion (positive, negative, or neutral), this process is named Sentiment Analysis.

Data and Analysis

This section discusses data collection, pre-processed and vectorized the text to apply ML algorithms (Decision Tree, Random Forest, Naïve Bayes and Logistic Regression) to the corpus dataset.

Data Collection

The dataset contained four years of tweets data. These tweets collected from Tweedy library full archive and sandbox. The tweets were about Saudi Arabian tourism. The dataset was collected between January 2017 and February 2021, with 273 K of tweets. In data collection, the aim was to retrieve all relevant tweets to a specific keyword. The list of keywords or hashtags used to collect tweets were selected based on trending hashtags related to KSA tourism. Additionally, the retrieved tweets filtered based on the tweets' geographical location to verify that the tweets were generated from Saudi Arabia. Moreover, some filters were added such as the language as the aim of this research to analyze Arabic tweets.

The data was gathered using Twitter's API and saved as a .csv file. It included the following information: User name, time, date, user followers, who the user was following, location and the textual comment. The tweet was filtered if it contained keywords in the Arabic language such as ("#السياحة", "#السياحة_السعودية", "#عيش_السعودية"، "#زور_السعودية"، "#حولك") (Tourism-Saudi-Arabia, Live-Saudi-Arabia, Visit-Saudi-Arabia, around you). Initially the data contained 273800 records, the tweet length was between 3 and 575 characters.

Pre-Processing

The data pre-processing included tweets dataset and labelled data (corpus)

Tweets Dataset

Pre-processing steps had been applied to the dataset of tweets and the labelled data (corpus):

- Elimination of duplicates rows
- Extraction of Arabic text by removing non-Arabic text, symbols (#,), punctuations, emojis
- Functions from araby library had been applied to normalize the morphology of the text

- araby. strip_tatweel (text)
- araby.normalize_ligature (text)
- Elimination of null values

After that the Lemmatisation was implemented and testing, the results showed that the performance model was decreased. Therefore, the text without Lemmatisation was used to reach better performance.

To visualize the data, the word cloud representation was used. Word cloud is visual representation of the most used keywords. In general, words are displayed in sizes and fonts that are more visible the more used or popular the words are. After pre-processing, the word cloud of tweets is represented in Fig. 2.

Corpus

The Ara Senti corpus was used as the labelled data in this research study (AL-Twairesh *et al*., 2017). It is annotated corpus of Arabic tweets for sentiment analysis. Modern Standard Arabic and the Saudi dialect were used in tweets.

Fig. 2: Word cloud after pre-processing

Fig. :3 Word cloud after pre-processing of corpus

The corpus was annotated manually by three annotators. It contained 17,573 tweets labelled with four different opinion labels: Positive, negative, neutral and mixed. Only positive, negative and neutral labels were used to represent three classes. Before choosing the Ara Senti corpus, other labeled data corpora were tested. These labeled data are:

- NLP-for-dialectDetection-TopicDetection-SentimentAnalysis-master: Dataset of Arabic tweets written with different dialects; Algeria, Egypt, Lebanon, Tunisia and Morocco, labelled with 3 classes: Positive, Negative and Neutral
- MASC Corpus: Dataset written in Arabic with different dialects and treats different subjects, The data labels with two classes: Positive and negative
- ASTD: Dataset from Egypt and labelled with four classes; Positive, Negative, Neutral and Mixed
- AJGT: Arabic Jordanian General Tweets (AJGT) Corpus labelled with two classes; Positive and Negative

After these corpora were tested, the best model performance was found through using Ara Senti corpus. Also, Ara Senti corpus was fit the tweet dataset in this study since it is written in MSA (Modern Standard Arabic words) and the Saudi dialect. The word cloud of the corpus text after pre-processing is presented as shown in Fig. 3.

Implementation and Results

The steps for implementing text classification task are presented in Fig. 4. These implementation steps were used for tweets dataset and corpus The first step was text pre-processing using specific tools for Arabic text, then tokenization and features extractions methods. Different features extraction methods were tested: Bag of words (unigram and bigrams models) and TF-IFD. The corpus was split into training set and testing set, 80% for train data and 20% for test data.

Different Machine learning models were tested. These models were logistic regression, decision tree, random forest and naïve bayes. After trained and tested the models, trainable model was used to categorize the tweets to three class positive, negative and natural. Python and Apache spark were used to implement this model. Python is a programming language placed under a free license. Python libraries used in this research study included Tweepy, MLlib, PyArabic, Qalsadi Arabic Morphological Analyzer and the NLTK (natural language toolkit) as well as other standard libraries (pandas, regex, string and matplotlib). Apache Spark as a fast data processing tool dedicated to big data was also used in this study. The Apache Spark's machine learning library (MLlib) was used to perform ML models, tokenization and feature extraction. Machine learning pipelines were made up iterative steps that were used to train the model to help automate ML workflows. The Pipelines included tokenization, feature extraction and modelling.

For TF-IDF representation, two functions were used: Count Vectorizer and Hashing TF from the MLlib library. The term frequency vector produced using either function. The training corpus determined the size of the vector produced by Count Vectorizer. The document was transformed to fixed-size vectors for the Hashing TF function. The default dimension of feature was 262,144. A Hash Function (Murmur Hash 3) was used to mapped terms to indices. The implementation of the two models (logistic regression and naïve bayes) were with Hashing TF + IDF, Bag of word model (unigrams) and Bi-grams model. The other two models (random forest and decision tree) were with Hashing TF + IDF and Bag of word model (unigrams). An example of implementing four models with the TF-IDF representation using Count Vectorizer function to test the training dataset are shown in Table 1, 2, 3, 4 respectively. In these tables: (1) Words is the output of tokenization function-list of tokens, (2) CV is the output of the Count Vectorizer function-the counts of token of the document over the vocabulary, (3) Features are the output of the IDF function, (4) Raw prediction is the raw output of the logistic regression classifier, (5) Probability is the result of applying the logistic function to Raw Prediction array and (6) Prediction represents the predicted class corresponding to the maximum value of the probability.

Evaluation the Models

To choose which model provide better prediction results, the comparison was made through evaluation metrics. The results in Table 5 demonstrated that the Logistic Regression and Naïve Bayes achieved better performance with around 86%, Logistic regression model combined with TF-IDF representation and Naïve Bayes with the bag of words representation. Applied adding words context by extracting bi-grams, the model's efficiency was not improved. Random Forest and decision Tree are robust algorithms; however, these algorithms are not a best choice for high-dimensional sparse matrix. For Random Forest model the F measure and the precision did not exceed the values of 48 and 69% respectively. For Decision Tree the precision was about 78%. Using Hashing TF or Connectorized function for features extracting with TF-IDF method, did not impact the metrics.

Prediction Tweets Labels

Based on evaluation metrics, the Logistic Regression model was adopted with TF-IDF method using Connectorized. The elaborated classifier was used to predict classes on the collected dataset using the created pipeline to transform data and predict the classes. The predicted classes are distributed in Table 6. The tweets dataset was classified as 22 k of the tweets are Negative, 140 k as Neutral and 46 k as Positive.

Table 1: The results from applying LR and CV

	sentence	Label	Words	CV	Features	Raw prediction	Probability	Prediction
1351	انا صدق الخبر فهو تطور ممتاز لأنهم اعتادو...	2.0	(اذا, صدق, الخبر, فهو, تطور, ممتاز, جدا, لأأنهم...)	(1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, …	(1.4692065 1122138,0.0, 0.0, 0.0, 0.0, 2.10105…	[0.1604417472 874546, -2.251 864916472966, 2.091….	[0.1255220961 20663656, 0.01122084492 6712794, 0….	2.0
199	تذكر جيدا وستعرف مدى صحة كلامى	0.0	(تذكر, جيدا, وستعرف, مدى, صحة , كلامى)	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, ….	(1.4692065112 2138,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[0.267505296952 68394, -0.14 8727 87363290346, -0….	[0.42751397184 88128, 0.281956 7808919869, 0.290….	0.0
849	أغلق مؤشر سوق الأسهم السعودية اليوم مرتفعا....	1.0	(أغلق, مؤشر, سوق, الأسهم, السعودية, اليوم, مرتفعا...)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	[-1.448537011 712598, 2.575 470290890736, -1.126….	[0.017151451609 00047, 0.9591908 980450166, 0.02….	1.0
1256	اتفق مسلسل رائع	2.0	(اتفق, مسلسل, رائع)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.27369202…	[-0.550085042 2368379, -0.57 34142236670552, 1.1….	[0.136832912654 45136, 0.13367 766063562678, 0.7….	2.0
9062	يوم مانى مقتنع فيه للأسف...	0.0	(يوم, مانى, مقتنع, فيه, للأسف...)	(2.0, 0.0, 0.0, 0.0, 0.0, 0 .0, 0.0, 0.0, 0.0, …	(2.9384130224 4276,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[-0.550597236 7411923, -0.31 985413959437 414, -0….	[0.532891532751 0582, 0.223154 98361195313, 0.24….	0.0

DOI: 10.3844/jcssp.2022.215.226

Table 2: The results from applying Decision Tree and CV

	Sentence	Label	Words	CV	Features	Raw prediction	Probability	Prediction
1351	انا صدق الخبر فهو تطور ممتاز لأنهم اعتادو...	2.0	(اذا, صدق, الخبر, فهو, تطور, ممتاز, جدا, لأنهم...).	(1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, …	(1.4692065112213 8,0.0, 0.0, 0.0, 0.0, 2.10105…	[11.0, 11.0, 731.0]	[0.01460823373173 9707,0.0146082337 31739707, 0….	2.0
199	تذكر جيدا وستعرف مدى صحة كلامى	0.0	(تذكر, جيدا, وستعرف, مدى, صحة , كلامى)	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, ….	(1.469206511221 38,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[1819.0, 3144.0, 931.0]	[0.3086189345096 709, 0.53342382 08 347472, 0.157…	1.0
849	أغلق مؤشر سوق الأسهم السعودية اليوم مرتفعا....	1.0	(أغلق, مؤشر, سوق, الأسهم, السعودية, اليوم, مرتفعا...)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	[1819.0, 3144.0, 931.0]	[0.3086189345096 709, 0.53342382 08 347472, 0.157…	1.0
1256	اتفق مسلسل رائع	2.0	(اتفق, مسلسل, رائع)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.27 369202…	[1.0, 0.0, 926.0]	[0.001078748651 5641855, 0.0, 0.9989 21251348 4358]	2.0
9062	يوم مانى مقتنع فيه للأسف...	0.0	(يوم, مانى, مقتنع, فيه, للأسف...)	(2.0, 0.0, 0.0, 0.0, 0.0, 0 .0, 0.0, 0.0, 0.0, …	(2.93841302244 276,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[1819.0, 3144.0, 931.0]	[0.30861893450967 09,0.5334238208 347 472, 0.157…	1.0

Table 3: The results from applying Random Forest and CV

	Sentence	Label	Words	CV	Features	Raw prediction	Probability	prediction
1351	انا صدق الخبر فهو تطور ممتاز لأنهم اعتادو...	2.0	(اذا, صدق, الخبر, فهو, تطور, ممتاز, جدا, لأنهم...).	(1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, …	(1.46920651122138, 0.0, 0.0, 0.0, 0.0, 2.10105…	[5.04294957259636,6. 77653 2273375356, 8.180518….	[0.252147478629818, 0.3388 2661366 876776,0.409….	2.0
199	تذكر جيدا وستعرف مدى صحة كلامى	0.0	(تذكر, جيدا, وستعرف, مدى, صحة , كلامى)	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, ….	(1.46920651122 138,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[5.6820002203795, 7.324173 192356 255, 6.9938265….	[0.28410001101897 5, 0.3662 0865961 78128, 0.3496….	1.0
849	أغلق مؤشر سوق الأسهم السعودية اليوم مرتفعا....	1.0	(أغلق, مؤشر, سوق, الأسهم, السعودية, اليوم, مرتفعا...)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	[4.523910234948667, 8.9787 07484574 077, 6.49738….	[0.226195511747433 33, 0.44893537422 870383, 0.3…	1.0
1256	اتفق مسلسل رائع	2.0	(اتفق, مسلسل, رائع)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.27 369202…	[5.170769016399003, 7.4447 20282270858 5, 7.3845….	[0.25853845081995 014, 0.37223601411 35429, 0.36…	1.0
9062	يوم مانى مقتنع فيه للأسف...	0.0	(يوم, مانى, مقتنع, فيه, للأسف...)	(2.0, 0.0, 0.0, 0.0, 0.0, 0 .0, 0.0, 0.0, 0.0, …	(2.938413022442 76,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[5.1625771679431605, 7.213 27789222602, 7.62414….	[0.2581288583971 5805, 0.360663894 611301, 0.381…	2.0

Table 4: The results from applying NB and CV

	sentence	Label	Words	CV	Features	Raw prediction	Probability	prediction
1351	انا صدق الخبر فهو تطور ممتاز لأنهم اعتادو...	2.0	(اذا, صدق, الخبر, فهو, تطور, ممتاز, جدا, لأنهم...).	(1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, …	(1.469206511221 38,0.0, 0.0, 0.0, 0.0, 2.10105…	[-733.65365622 4691,-899.233336 187621, -713.7….	[2.3453630630418 21e-09, 2.8831601 46687965e-81, ….	2.0
199	تذكر جيدا وستعرف مدى صحة كلامى	0.0	(تذكر, جيدا, وستعرف, مدى, صحة , كلامى)	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, ….	(1.469206511221 38,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[-320.54620561 074 364, -356.08 314 324935753, -32….	[0.99988067698310 92, 3.6851252682 248793e-16, 0….	0.0
849	أغلق مؤشر سوق الأسهم السعودية اليوم مرتفعا....	1.0	(أغلق, مؤشر, سوق, الأسهم, السعودية, اليوم, مرتفعا...)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …	[-861.466281740 9941, 602.27970 933 45001, -847….	[2.73339137964805e -113, 1.0, 2.363125 197662704….	1.0
1256	اتفق مسلسل رائع	2.0	(اتفق, مسلسل, رائع)	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, …	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2. 27369202…	[-135.650169372 60435, -158.7346 919 4287765, -10….	[4.334938048756837e -14, 4.08791752 93234e-24, 0….	2.0
9062	يوم مانى مقتنع فيه للأسف...	0.0	(يوم, مانى, مقتنع, فيه, للأسف...)	(2.0, 0.0, 0.0, 0.0, 0.0, 0 .0, 0.0, 0.0, 0.0, …	(2.9384130224427 6,0.0, 0.0, 0.0, 0.0, 0.0, 0…	[-87.596717143 3118, -104.38498 600 651343, -97.1….	[0.9999261109461762, 5.115 8214815757 026e-08, 7….	0.0

Table 5: The comparison of the four models

Model	Features extraction	F1 score	Precision	Recall	
Logistic Regression	Bag of words	Uni-gram	81,91%	82,97%	82,19%
	TF-IDF	Count Vectorizer - Unigram	**86,38%**	**86,63%**	**86,52%**
		Count Vectorizer - Ngram (n = 2)	64,43%	71,06%	65,34%
		Hashing TF	85,97%	86,17%	86,11%
Decision Tree	Bag of words	Uni-gram	63,98%	78,14%	66,59%
	TF-IDF	Count Vectorizer	63,98%	78,14%	66,59%
		Hashing TF	63,62%	78,34%	66,41%
Random Forest	Bag of words	Uni-gram	43,97%	65,87%	52,01%
	TF-IDF	Count Vectorizer	47,05%	68,42%	55,38%
		Hashing TF	43,54%	68,25%	52,24%
Naive Bayes	Bag of words	Uni-gram	**85,48%**	**86,78%**	**85,65%**
		N-gram (n=2)	72,53%	77,11%	73,23%
	TF-IDF	Count Vectorizer	81,76%	82,38%	81,77%
		Hashing TF	81,51%	82,14%	81,63%

Table 6: Prediction

Opinion	Tweets numbers
Negative	22 407
Neutral	140 569
Positive	46 681

Table 7: Most visited places

	Location	Negative	Neutral	Positive
0	الرياض	393	4273	974
1	العلا	261	3535	473
2	حائل	204	2750	560
3	الطائف	59	2877	453
4	تبوك	211	2706	178

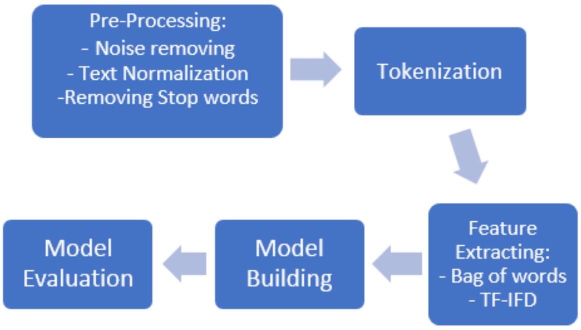

Fig. 4: Implementation Process.

Search for Location

After predicting the class of each tweet, the next step was searching for the location among tweet text to compile the most visited places in Saudi Arabia with their sentiment analysis. The searching for location process is as follows:

1. Creating a list of keywords (places) that was used to search for places in the text. This list contained different forms of names. For example, search for الطايف (Tayif) and الطائف (Taif)
2. A database with normalized place names was created. For example, for the word الطايف (Taef) the normalized form is الطائف (Taif)
3. A local function named "location" was created and introduced some rules to find places on the text as follows:

   - If the word found like وادي on the text, extract the وادي word and the next word, same for (Island)جزر , قرية (Village)
   For instance, the word وادي (Wadi) indicates the existence of a valid location وادي حنيفة. (Wadi Hanifa)
   For words like مهد (Mahd) and بدر (Badr), it should be followed by a specific word to consider them as a correct location and extract it to: مهد الذهب, بدر الجنوب (Badr AlJanub, Mahd Alhahab)

4. After extracting the different forms of place names, a local function named "normalize_location" was used to normalize these place names. This function returned for each name its normalized form if it existed
5. The most visited places were selected; the 50 most visited places were considered in this study
6. For each location, the number of tweets was calculated per sentiment class

The results demonstrated five most visited places in Saudi Arabia, which are Riyadh, Alula, Hail, Taif and Tabuk respectively. These places with the correspondent predicted opinion are shown in Table 3.

Table 7 shows the top five visited places on Saudi Arabia are: Riyadh, Alula, Hail, Taif and Tabuk with sentiment analysis of each place. For example, Riyadh had 393 negative tweets, 4273 neutral tweets and 974 positive tweets. This analysis result is of great interest to a variety of stakeholders, including tourism organizations, ministries of tourism, travel companies.

Conclusion

The field of sentiment analysis is explored, like all other fields of natural language processing. This field has supported with a major evolution with the availability of data collected from social media platforms. In this research study, the Pre-processing techniques on Arabic text, different feature extraction techniques and Several Machine Learning models were implemented. MLlib Apache Spark's scalable machine learning library on python and ML algorithms were used on large tweets data. The Decision Tree, Random Forest, Logistic Regression and Naïve Bayes were applied with bag-of-words, bigram model and TF-IDF. The evaluation of the model was through metrics such as precision, recall and F1-score. The results showed the efficiency of Logistic Regression and Naïve Bayes on Arabic text classification. The novelty of this research is to explore the tourism industry which other researchers have not done with data analytics Twitter.

In the future work, deep learning approaches can be used with different architectures (Long Short-Term Memory, Neural Networks, Recurrent Neural Network (RNN) etc.) to perform the text classification and expanding the current research for mixed languages (Arabic and English text).

Author's Contributions

Wala Awadh Alasmari: Contributed to literature review, implementing the proposed algorithms, analyzing the results and writing of the manuscript

Hoda Ahmed Abdelhafez: Contributed to conceptualization, co-analyzing the results, editing and reviewing the manuscript.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. Journal of Travel Research, 58(2), 175-191. doi.org/10.1177/0047287517747753

Aldayel, H. K., & Azmi, A. M. (2016). Arabic tweets sentiment analysis–a hybrid scheme. Journal of Information Science, 42(6), 782-797. doi.org/10.1177/0165551515610513

Alhajji, M., Al Khalifah, A., Aljubran, M., & Alkhalifah, M. (2020). Sentiment analysis of tweets in Saudi Arabia regarding governmental preventive measures to contain COVID-19. doi.org/10.20944/preprints202004.0031.v1

Alomari, E., Katib, I., Albeshri, A., & Mehmood, R. (2021). COVID-19: Detecting government pandemic measures and public concerns from Twitter arabic data using distributed machine learning. International Journal of Environmental Research and Public Health, 18(1), 282. doi.org/10.3390/ijerph18010282

Alosaimi, S., Alharthi, M., Alghamdi, K., Alsubait, T., & Alqurashi, T. (2020). Sentiment Analysis of Arabic Reviews for Saudi Hotels Using Unsupervised Machine Learning. Journal of Computer Science, 16(9), 1258-1267. doi.org/10.3844/jcssp.2020.1258.1267

AL-Rubaiee, H. S., Qiu, R., Alomar, K., & Li, D. (2016). Sentiment analysis of Arabic tweets in e-learning. doi.org/10.3844/jcssp.2016.553.563

Alruily, M., & Shahin, O. R. (2020). Sentiment Analysis of Twitter Data for Saudi Universities. International Journal of Machine Learning and Computing, 10(1). doi.org/10.18178/ijmlc.2020.10.1.892

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., & Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Computer Science, 117, 63-72. doi.org/10.1016/j.procs.2017.10.094

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Boujou, E., Chataoui, H., Mekki, A. E., Benjelloun, S., Chairi, I., & Berrada, I. (2021). An open access NLP dataset for Arabic dialects: Data collection, labeling and model construction. arXiv preprint arXiv:2102.11000. doi.org/10.48550/arXiv.2102.11000

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. doi.org/10.1023/a:1010933404324

Campan, A., Atnafu, T., Truta, T. M., & Nolan, J. (2018, December). Is data collection through twitter streaming api useful for academic research?. In 2018 IEEE international conference on big data (Big Data) (pp. 3638-3643). IEEE. doi.org/10.1109/bigdata.2018.8621898

Chen, W., Xu, Z., Zheng, X., Yu, Q., & Luo, Y. (2020). Research on sentiment classification of online travel review text. Applied Sciences, 10(15), 5275. doi.org/10.3390/app10155275

Duwairi, R. M. (2015, April). Sentiment analysis for dialectical Arabic. In 2015 6th international conference on information and communication systems (ICICS) (pp. 166-170). IEEE. doi.org/10.1109/iacs.2015.7103221

Duwairi, R., & El-Orfali, M. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. Journal of Information Science, 40(4), 501-513. doi.org/10.1177/0165551514534143

El Alaoui, I., & Gahi, Y. (2019). The impact of big data quality on sentiment analysis approaches. Procedia Computer Science, 160, 803-810. doi.org/10.1016/j.procs.2019.11.007

Freihat, A. A., Abbas, M., Bella, G., & Giunchiglia, F. (2018). Towards an optimal solution to lemmatization in arabic. Procedia computer science, 142, 132-140. doi.org/10.1016/j.procs.2018.10.468

Jo, T. (2021). Machine Learning Foundations: Supervised, Unsupervised and Advanced Learning. Springer Nature.

Kubat, M. (2017). *An Introduction to Machine Learning* 2nd ed., Springer International Publishing. ISBN: 9783319639130. PP. 348

Lin, Y. (2021). 10 Twitter Statistics Every Marketer Should Know in. Oberlo. https://www.oberlo.com/blog/twitter-statistics

DOI: 10.3844/jcssp.2022.215.226

Manorathna, R. (2020). Random forests--An ensemble of decision trees (This is how decision trees are combined to make a random forest). https://towardsdatascience.com/random-forests-an-ensemble-of-decision-trees-37a003084c6c (Accessed on Jan 31, 2021)

MAS (2012) Tourism Statistics 201. Riyadh: Tourism Information and Research Centre, SCTA.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. Information & Management, 54(6), 771-785. doi.org/10.1016/j.im.2016.11.011

MLlib. (2021). Main Guide - Spark 3.1.1 Documentation. http://spark.apache.org/docs/latest/ml-guide.html

Mubarak, H., & Darwish, K. (2014, October). Using Twitter to collect a multi-dialectal corpus of Arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP) (pp. 1-7). https://aclanthology.org/W14-3601.pdf

Ntaliakouras, N., Vonitsanos, G., Kanavos, A., & Dritsas, E. (2019, July). An apache spark methodology for forecasting tourism demand in Greece. In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA) (pp. 1-5). IEEE. doi.org/10.1109/IISA.2019.8900739

NTP. (2021). National Transformation program. https://www.vision2030.gov.sa/en/programs/NTP

Omari, M. A., & Al-Hajj, M. (2020). Classifiers for Arabic NLP: Survey. International Journal of Computational Complexity and Intelligent Algorithms, 1(3), 231-258. doi.org/10.1504/ijccia.2020.105538

Omnicore, Twitter by the Numbers (2021): Stats, Demographics & Fun Facts. (2021). https://www.omnicoreagency.com/twitter-statistics

Rajput, A. (2020). Natural language processing, sentiment analysis and clinical analytics. In Innovation in Health Informatics (pp. 79-97). Academic Press. doi.org/10.1016/B978-0-12-819043-2.00003-4

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). Aravec: A set of Arabic word embedding models for use in arabic nlp. Procedia Computer Science, 117, 256-265. doi.org/10.1016/j.procs.2017.10.117

Statista, Twitter: Most users by country (2021). https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A survey on distributed machine learning. ACM Computing Surveys (CSUR), 53(2), 1-33. doi.org/10.1145/3377454

Yuliana, Y., & Erlangga, E. (2017, October). Analysis Of Data Mining Methods Naive Bayes Classifier (NBC). In International Conference on Engineering and Technology Development (ICETD). http://artikel.ubl.ac.id/index.php/icetd/article/view/1030

Zerrouki, T. (2010). Tashaphyne, Arabic light stemmer. https://pypi.python.org/pypi/Tashaphyne/0.2

Zerrouki, T. (2021). PyArabic (An Arabic language library for Python). https://pypi.python.org/pypi/pyarabic

Zhiqiang, D., & Changguo, X. (2016, June). Research on Intelligent Tourism Application Based on Big Data. In 7th International Conference on Education, Management, Information and Computer Science (ICEMC 2017) (pp. 493-496). Atlantis Press. doi.org/10.2991/icemc-17.2017.98