Real-time Twitter Sentiment Analysis for Moroccan Universities using Machine Learning and Big Data Technologies

https://doi.org/10.3991/ijet.v18i05.35959

Imane Lasri(✉), Anouar Riadsolh, Mourad Elbelkacemi
Mohammed V University in Rabat, Rabat, Morocco
imane_lasri@um5.ac.ma

Abstract—In recent years, sentiment analysis (SA) has raised the interest of researchers in several domains, including higher education. It can be applied to measure the quality of the services supplied by the higher education institution and construct a university ranking mechanism from social media like Twitter. Hence, this study presents a novel system for Twitter sentiment prediction on Moroccan public universities in real-time. It consists of two phases: offline sentiment analysis phase and real-time prediction phase. In the offline phase, the collected French tweets about twelve Moroccan universities were classified according to their sentiment into 'positive', 'negative', or 'neutral' using six machine learning algorithms (random forest, multinomial Naive Bayes classifier, logistic regression, decision tree, linear support vector classifier, and extreme gradient boosting) with the term frequency-inverse document frequency (TF-IDF) and count vectorizer feature extraction techniques. The results reveal that random forest classifier coupled with TF-IDF has obtained the best test accuracy of 98%. This model was then applied on real-time tweets. The real-time prediction pipeline comprises Twitter streaming API for data collection, Apache Kafka for data ingestion, Apache Spark for real-time sentiment analysis, Elasticsearch for real-time data exploration, and Kibana for data visualization. The obtained results can be used by the Ministry of higher education, scientific research, and innovation of Morocco for the decision-making process.

Keywords—higher education, sentiment analysis, machine learning, big data, Twitter

1 Introduction

Nowadays, the growth of social media websites offers an opportunity to people to communicate their sentiments, opinions, and ideas concerning a given topic. Twitter is considered one of the widely-used websites, with 435 million monthly active users [1]. Data extracted from this media has motivated researchers to explore public opinions. Hence, Twitter sentiment analysis has become an attractive research topic. It identifies whether sentiment polarity from text data is positive, neutral, or negative by analyzing the opinion to help improve the decision-making process.

Numerous studies have shown that Twitter sentiment analysis is more efficient when used in certain domains including healthcare [2–3], banking sector [4], marketing [5–6], tourism [7], and politics [8]. Moreover, sentiment analysis is used in the education field to improve the quality of the educational institution's services [9–10], enhance the learning process by analyzing students' feelings [11] and orientation [12], and extract useful information about the teaching methodology of a teacher. In particular, recent studies have shifted their focus to sentiment analysis to detect the strengths and weaknesses of courses in higher education by analyzing students' online opinions [13] and measuring specific university indicators such as university reputation from social media for constructing a ranking mechanism [14]. Analyzing tweets about universities can be a vital complementary source for policymakers in higher education environments for budget allocation and university evaluation.

Sentiment analysis can be carried out using machine learning and lexicon-based. Lexicon-based approaches rely on the assumption that the text's semantic orientation is related to the polarity of words and phrases that occur in it, and machine learning approaches are based on models that are constructed from labeled examples of sentences. However, real-time sentiment analysis necessitates powerful big data technologies, such as Apache Spark [15] and Apache Kafka [16] for data processing. Then, Elasticsearch [17] and Kibana for real-time storing and visualizing the vast amount of data.

To the best of our knowledge, no previous studies have been conducted for analyzing Moroccan universities-related tweets. Hence, this drives us to introduce a novel system for Twitter sentiment analysis about Moroccan public universities in real-time with machine learning and big data. The predicted results will increase the university ranking systems in Morocco by including people's sentiments (positive, negative, and neutral) concerning such academic institutions.

The main contributions of this study are as follows:

- Introducing a real-time system for predicting the sentiment for Moroccan public universities from Twitter using big data technologies.
- Collecting tweets data in French about twelve Moroccan public universities then pre-processing and labeling them.
- Conduct comparative studies on the performance of different machine learning classifiers with term frequency-inverse document frequency (TF-IDF) and count vectorizer.
- Selecting and applying the best model with the highest test accuracy on a new unlabeled tweet collected in real-time to predict people's opinions regarding Moroccan public universities.

The remainder of this paper is categorized as follows. Section 2 presents the related works. Then, we describe the proposed methodology in Section 3. The experiment setup and experimental results are discussed in Section 4. Finally, Section 5 presents the conclusion and future work.

2 Related work

The vast development of social media usage has presented an opportunity for every organization, company, and institution to get user opinions, comments, and reviews on specific issues. As of January 2022, Twitter has 435 million monthly active users [1]. Thus, Twitter information is valuable for sentiment analysis. Accordingly, numerous prior studies have focused on analyzing opinions from Twitter data in various domains including movies, tourism, and, more lately, education.

Along with the great success of machine learning in many application domains, machine learning algorithms are also applied in sentiment analysis. Neethu and Rajasree [18] analyzed Twitter posts about electronic products using machine learning algorithms. Goel et al. [19] classified movie reviews from Twitter in real-time using Naive Bayes [20]. The authors also explained that the Naive Bayes accuracy could be improved by using it together with SentiWordNet. Chikersal et al. [21] analyzed tweets using supervised learning with a rule-based classifier. The dataset used for training Support Vector Machines (SVM) [22] comprised 9418 tweets. Coletta et al. [23] used a different method that combines the SVM with a cluster ensemble to analyze Twitter data. It has shown better test accuracy than a stand-alone SVM. Later, SVM and sentiment classification algorithm (SCA) were used by Huq et al. [24] on Twitter Data to find sentiment analysis. The performance of SVM and SCA was compared, and then SCA is found to be better than SVM.

Sentiment analysis has also attracted researchers' attention in the education domain recently. Many studies have applied sentiment analysis to analyze students' attitudes toward various aspects using machine learning and deep learning. Altrabsheh et al. [25] built a student response system to collect feedback from Twitter, clickers, and mobile phones in real-time then they compared different sentiment analysis techniques like SVM, Naive Bayes, and Max Entropy. Mujahid et al. [26] analyzed the sentiments of people regarding distance education during COVID-19 using deep learning and machine learning techniques on a dataset that contains 17,155 tweets about e-learning. Waheeb et al. [27] proposed a method to analyze sentiments of e-learning from Twitter to identify COVID-19 fake news using the Extreme Learning Machine-Autoencoder (ELM-AE) with LSTM.

In the higher education context, many studies have been applied to reviews about universities and colleges using machine learning techniques. Balachandran and Kirupananda [28] examine the users' feedback on Twitter and Facebook about Sri Lanka universities with the StanfordCoreNLP library. Alruily and Shahin [29] proposed a sentiment analysis approach for analyzing tweets for Saudi universities using SVM, Naive Bayes, K-Nearest Neighbors classifier (KNN), Random Forest [30], Stochastic Gradient Descent (SGD), Sequential Minimal Optimization (SMO), Multilayer Perceptrons, and Sentiment Score Calculation (SSC). The collected dataset contains 5882 positive comments and 5882 negative comments about 22 Saudi universities. The best test accuracy is achieved using SVM. AL-Rubaiee et al. [31] introduced a framework for analyzing Arabic tweets regarding King Abdul-Aziz University students' attitudes toward distance learning using Naive Bayes and SVM.

3 Proposed methodology

The overall architecture of the proposed system for real-time sentiment prediction about Moroccan universities-related tweets is depicted in Figure 1. It comprises two major components: an offline sentiment analysis model using machine learning algorithms and a real-time sentiment analysis pipeline.

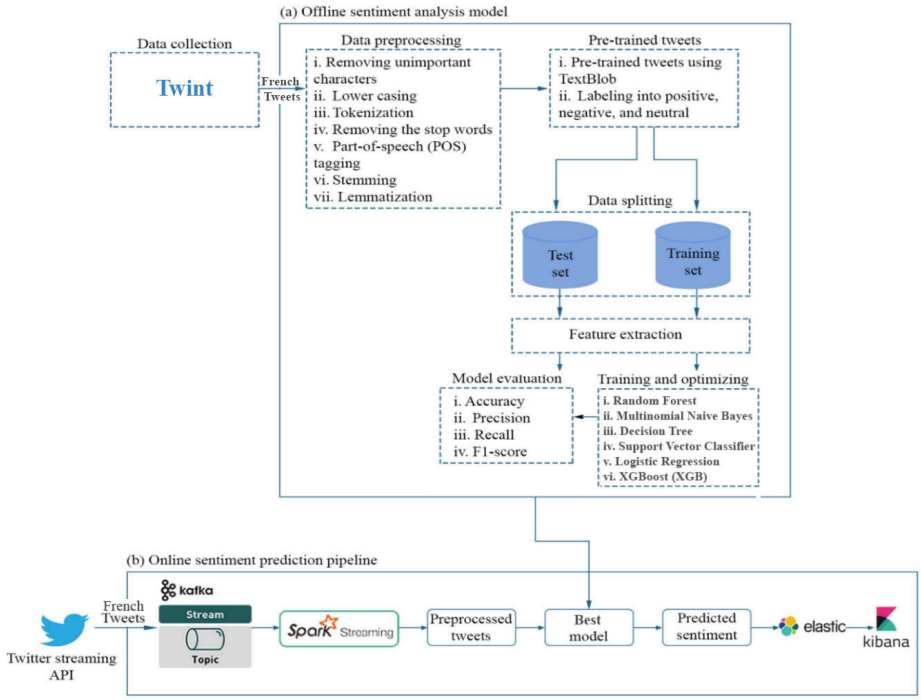

Fig. 1. Real-time sentiment analysis architecture

3.1 Offline sentiment analysis phase

We developed a sentiment analysis model using six machine learning algorithms with different feature extraction techniques to be applied for real-time sentiment prediction. We used a collected French tweets dataset about twelve Moroccan universities to train the machine learning models. Figure 1 illustrates the offline sentiment analysis stages, comprising data collection, data pre-processing, pre-trained tweets, data splitting, feature extraction, sentiment analysis model, and model performance evaluation.

Data collection. In the data collection stage of the offline sentiment analysis model, we collected 1798 historical French tweets related to 12 Moroccan public universities between June 11, 2009, and May 24, 2022. Since the Search API allows access to old data dating back up to one week before, we have used Twint [32] an open-source Python library designed to extract tweets based on keywords, dates, locations, and hashtags. The number of collected tweets of each university is shown in Table 1.

Table 1. Number of tweets for each Moroccan University

University	Tweet Count
Mohammed V University in Rabat	375
Sidi Mohamed Ben Abdellah University	240
Ibn Tofail University	220
Cadi Ayyad University	158
Hassan I University	154
Ibn Zohr University	118
Abdelmalek Essaadi University	113
Chouaib Doukkali University	108
Hassan II University	106
Mohamed First University	85
Moulay Ismail University	61
Sultan Moulay Slimane University	60

Data pre-processing. Data generated on social media sites are vast, noisy, distributed, heterogeneous, and unstructured. Hence, removing the noise from the Twitter data is essential before applying machine learning or deep learning algorithms. To achieve this, we must perform some form of text pre-processing using the following techniques.

1. *Removing URLs, numbers, usernames, special and non-ASCII characters:* is the first pre-processing step in which URLs, numbers, special characters, non-ASCII characters, and user references begin with an @ symbol are removed because they do not play any role in sentiment analysis.
2. *Lower casing:* in this second step, every uppercase letter is changed to its corresponding lowercase letter.
3. *Tokenization:* this pre-processing technique splits sentences into smaller units called tokens using white space characters as delimiters.
4. *Removing the stop words:* is a technique to remove the noise from the tweets. Stop words are frequent words in natural language with only grammatical roles and no semantic value such as (is, an, the, etc.).
5. *Part-of-speech (POS) tagging:* is the essential pre-processing step that assigns a word to its grammatical category to understand its role within the sentence and build parse trees. Traditional parts of speech are nouns, verbs, conjunctions, adverbs, etc. Part of speech tagging provides the contextual information that a lemmatizer needs.
6. *Stemming:* is a technique to eliminate infixes, suffixed, prefixes, and circumfixes from a word. For example: ‘studied’ or ‘studying’ to ‘study’.
7. *Lemmatization:* is a process of removing plurals and conjugations and returning root words depending on their meaning and context. We used NLTK [33] lemmatization that transforms a sentence to its base form like ‘students’ to ‘student’ or ‘lessons’ to ‘lesson’.

Table 2. Example of implementing the pre-processing steps on a tweet

University	Before Pre-Processing	After Pre-Processing
Mohammed V University in Rabat	L'Université Mohammed V de Rabat a été classée première au Maghreb par le Center for World University Rankings (CWUR), qui a dévoilé, lundi 25 avril, les résultats de son classement annuel des universités mondiales. https://t.co/qo615wwp2D #Maroc #Enseignementsupérieur @um5rabat	université mohammed v rabat a classée première maghreb center for world university rankings cwur a dévoilé lundi avril résultat classement annuel université mondiales

Pre-trained tweets. We employed a TextBlob [34], an open-source Python NLP library, which gives polarity: a float in between −1 and 1, in which −1 characterizes a negative sentiment, 0 characterizes a neutral sentiment, and 1 characterizes a positive sentiment. Then, subjectivity: a float that lies between 0 and 1 and quantifies the number of personal emotions, feelings, or opinions in the text. The polarity is used, in our study, to label the collected tweets to be passed to the proposed sentiment analysis model.

The result was: 969 neutral tweets, 721 positive tweets, and 108 negative tweets.

Data splitting. To train our proposed sentiment analysis model, we used a stratified data splitting with a 90:10 ratio, where 90% of the pre-trained dataset is used for training and 10% for the test. In the training set each input text is related to the accurate sentiment, which allows the model to perform the training process. Otherwise, the test set consists of unseen data for model evaluation.

Feature extraction. There are several techniques for extracting irrelevant features from different data types. In this study, the count vectorizer and the term frequency-inverse document frequency (TF-IDF) [35] techniques are used in the proposed approach to extract semantic features from the tweet.

The Count Vectorizer technique consists of counting the occurrence of a term in a text document and uses this value as its weight as integers to convert a collection of text documents to a vector of term/token counts and builds a vocabulary of terms.

The term frequency-inverse document frequency (TF-IDF) aims to extract features from the text. The *TF* measures the number of a term *t* in document *d*, and *IDF* provides the term's importance. The *TF* and *IDF* are defined in Eq. (1) and (2).

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document\ d}{Total\ number\ terms\ in\ a\ document\ d} \tag{1}$$

$$IDF(t) = log_e \frac{Total\ number\ of\ documents}{Total\ number\ of\ documents\ with\ term\ t} \tag{2}$$

Then, the *TF−IDF* calculates the weight for a term *t* in a document *d* by multiplying the *TF* with *IDF* values for each term, as defined in Eq. (3).

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \tag{3}$$

TF−IDF is scored between 0 and 1. The higher values for less frequent words in the document corpus. The smaller value for significant words in the document corpus.

Classification models. In this research, the aim of the classification task is to assign unseen tweet to a sentiment class (positive, negative, or neutral) based on the training data. The machine learning algorithms used and their hyperparameters are described as follows:

1. *Decision tree [36]:* is a structure that includes a root node, branches, and leaf nodes, which represent an item in text classification. Each internal node represents a test for feature weights, each branch represents the feature, and each leaf node denotes a sentiment class. The model hyperparameters are criterion = entropy and splitter = best.
2. *Random forest [30]:* is an algorithm that consists of many decision trees. Each tree assigns a class to the input data, and the class with the highest occurrence is selected. The class of an unseen data is defined when it reaches the leaf of a tree. The number of estimators used with this model was set to 50.
3. *Multinomial Naive Bayes classifier [37]:* is an extension of a Naive Bayes algorithm, widely applied in text classification, in which the data are represented as word vector counts. It uses a multinomial distribution to calculate the probability of each class for a document and then returns the class with the highest probability. The Multinomial Naive Bayes equation is defined in Eq. (4), where w_i is the features vector, W is the frequency of a term i, C is the class.

$$P\left(W|C_k\right) = \frac{\left(\sum_{i=1}^{n} w_i\right)!}{\prod_{i=1}^{n} w_i\,!} \times \prod_{i=1}^{n} p_{k_i}^{w_i} \tag{4}$$

4. *Logistic regression [38]:* is a classification algorithm used to predict the probability of a categorical dependent variable. It uses the sigmoid function to assign the predicted values to probabilities based on a threshold value. Multinomial logistic regression is an extension of logistic regression used when the target variable has three or more values. It uses the softmax function defined in Eq. (5) instead of the sigmoid function.

$$\sigma(\vec{y})_i = \frac{e^{y_i}}{\sum_{j=1}^{C} e^{y_j}} \tag{5}$$

   Where C represents the number of classes, i represents the ith class, and y_i is the predicted score for the ith class. The model hypermapeters are multi_class = multinomial, solver = lbfgs, and $C = 5e1$.
5. *Linear support vector classifier (SVC):* was used in this research as it is similar to support vector machine (SVM) but with a rapid solver to avoid convergence problems for SVMs with linear kernel. For this experiment, C was set to 1 with dual equal to false.
6. *Extreme Gradient Boosting (XGBoost) [39]:* is a fast machine learning algorithm based on the gradient boosted decision trees technique. It functions in parallel tree boosting and is known for speed and scalability.

The grid search method with stratified 10-fold cross-validation (CV) was used to optimize the hyperparameters of the machine learning models and apply the best hyperparameters to predict the sentiment of a tweet.

Model evaluation. Four performance metrics are used for proposed model evaluation: accuracy, precision, recall, and $F1$-score. Equations (6), (7), (8), and (9) define the formulas of the evaluation metrics, where TP is true positive, TN is true negative, FP is false positive, and FN is a false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{9}$$

3.2 Real-time sentiment analysis pipeline

The real-time sentiment analysis pipeline predicts the sentiment of the Moroccan universities streaming tweets by performing real-time processing on the collected data using the deployed deep learning model. The tweet is extracted and published to a Kafka topic. A Spark streaming job consumed the message tweet from Kafka, performed sentiment analysis, and inserted the result into Elasticsearch. Then, the data stored on Elasticsearch are displayed on the Kibana dashboard. The architecture of the online prediction pipeline comprises five steps: data collection, data ingestion, real-time sentiment analysis, real-time data exploration, and data visualization.

Data collection using Twitter streaming API. Twitter streaming API is utilized to extract tweets from Twitter by using the Tweepy library, which enables Python to talk with the Twitter stage and utilizes its API via basic authentication ids such as consumer secret, consumer key, access secret, and access key. From this API many fields are scrapped like (text, source, retweets, language, user, location, etc..). The filter API helps for searching a stream of tweets that matches the hashtag. A total of 321 French tweets were collected in real-time for this study on 12 Moroccan public universities from May 25, 2022, to September 16, 2022.

Data ingestion using Apache Kafka. The streaming data from Twitter are ingested in real-time to Kafka topic. Data ingestion aims to import data from various sources such as social media sites, weblogs, relational database management systems (RDBMS), etc… to a target site for further processing. Apache Kafka [16] is a publish-subscribe distributed messaging system that has an important role in data ingestion in real-time. It is used to build real-time streaming data pipelines for transferring large volumes of

data across multiple applications. Kafka is run as a cluster on many servers named Kafka cluster, which consists of several brokers. A broker has a partition and each partition stores data streams with its keys, values, and timestamps in categories called topics. Kafka uses both push and pull mechanism, a producer pushes the message to the broker and consumers pulls the messages from the broker. In other words, the producer communicates with the Kafka brokers through the network for writing events on a topic, and consumers read these events. Kafka utilizes ZooKeeper [40] internally or externally to determine which broker is the leader of a given partition and topic and to do leader elections. Figure 2 illustrates the Apache Kafka architecture.

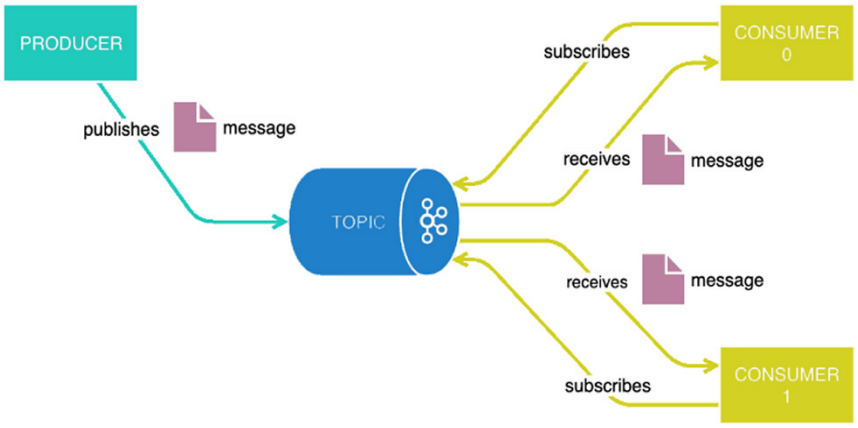

Fig. 2. Apache Kafka architecture

Parallel data processing with Apache Spark. Spark streaming and deep learning capabilities are used to carry out the prediction model for sentiment analysis. Apache Spark [15] is in-memory distributed open-source processing platform developed for fast computation. Resilient distributed dataset (RDD) is the primary Spark abstraction. It is a distributed collection of objects divided over a set of nodes of a cluster, with the ability to recover from faults occurrences in the system automatically. Spark follows master/slave architecture, which is a combination of SparkContext as the master of the Spark application, driver and a set of workers, that are acquired by Spark to run computations and store data for an application. Spark is written in Scala, but it also has APIs for Python, R, and Java. Figure 3 shows the Apache Spark architecture.

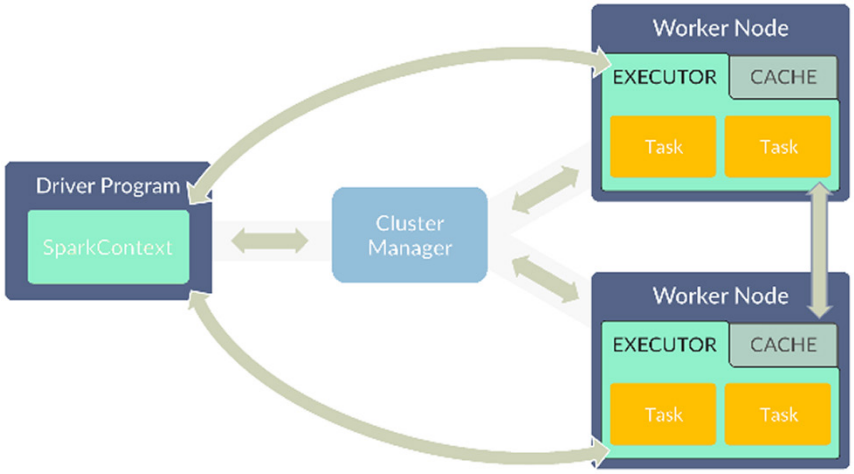

Fig. 3. Apache Spark architecture

Spark ecosystem has six components such as Spark Core, Spark MLlib, Spark SQL, Spark GraphX, SparkR, and Spark Streaming. Spark MLlib is a distributed machine learning framework over Spark Core. It has almost all popular machine learning algorithms including clustering, regression and classification. Spark Streaming gets live data streams, which are split into small batches called discretized streams (DStreams). They are a continuous sequence of RDDs in Spark memory that can be collected from many streaming data sources such as Flume, Kafka, and Elasticsearch. In our study, Spark streaming pre-processes the Moroccan university-related tweets to be fitted into the deep learning model in order to predict their corresponding sentiments.

Real-time data exploration with Elasticsearch. The online prediction results obtained using spark streaming API are emitted to Elasticsearch [17]. Elasticsearch is a real-time distributed engine built on top of Apache Lucene. Elasticsearch stores data in the form of a schema-less JSON (JavaScript Object Notation) documents. Each document has a set of keys with their corresponding values. Once the documents are indexed, Elasticsearch creates a data structure known as the inverted index designed to allow very fast search result retrieval. Data in Elasticsearch is distributed across multiple servers called Nodes. The collection of connected nodes is called a Cluster. Elasticsearch can support clients in many different languages such as PHP, Java, JavaScript, Python, Go, C#, and Ruby.

Data visualization with Kibana. The real-time processed data stored on Elasticsearch are displayed on the Kibana dashboard for analyzing people's sentiments regarding the Moroccan universities. Kibana is a data visualization layer built on top of Elasticsearch used to analyze and visualize the data in the form of a bar graph, line graph, pie charts, etc. Kibana works in sync with Elasticsearch and Logstash, which together forms the ELK stack. Kibana dashboards display the full pictures of Elasticsearch data and can be customized based on users' needs.

4 Experimental results

The performance evaluation of our proposed model is introduced in this section and explains the experimental setup and the results of the offline and real-time phases.

4.1 Experimental setup

The proposed model is written in the Python programming language and executed on Windows 10 Professional 64-bit operating system with Nvidia GTX 1070, Intel Core i7, 16 GB RAM, and a CPU of 3.20 GHz. In training, the dataset of the pre-trained tweets used in this study is splitted into training and test data set in a stratified 90–10 split ratio. We used the grid search method from the Sklearn library [41] to find the best hyperparameters values.

4.2 Results of the offline sentiment analysis phase

This section presents the obtained results of applying six machine learning models, including random forest (RF), multinomial Naive Bayes classifier (MNB), logistic regression (LR), decision tree (DT), linear support vector classifier (SVC), and extreme gradient boosting (XGBoost) with TF-IDF and count vectorizer. Each machine learning model is evaluated and presented in Figures 4 and 5 and Table 3.

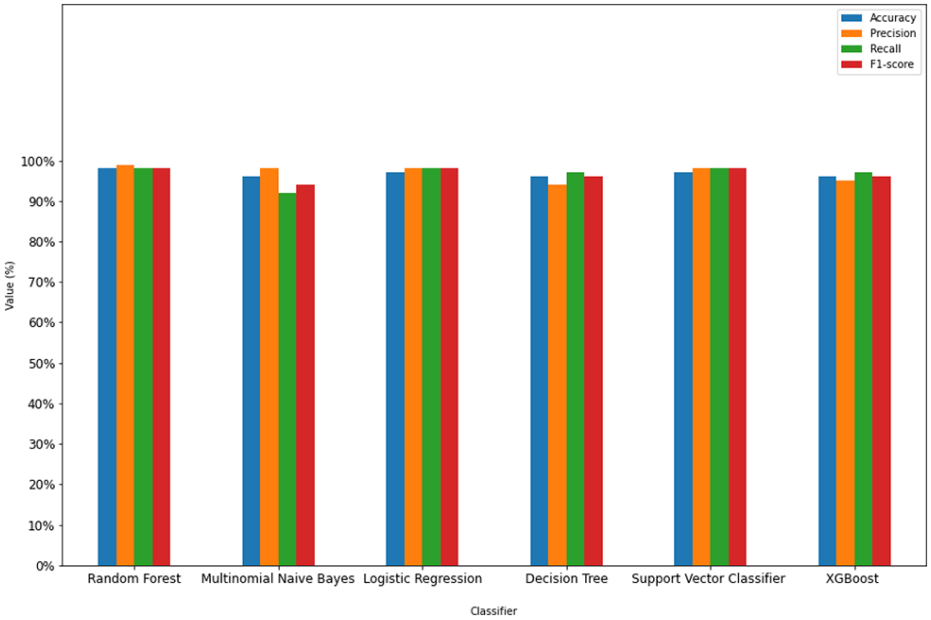

Fig. 4. Performance metrics of all machine learning algorithms with TF-IDF

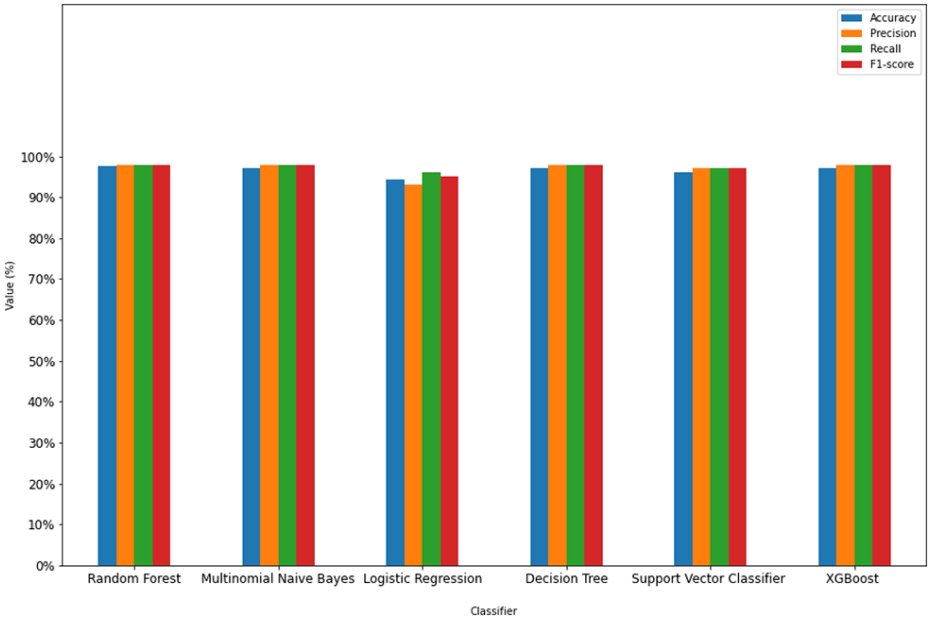

Fig. 5. Performance metrics of all machine learning algorithms with count vectorizer

Table 3. Performance results of six classifier with TF-IDF and count vectorizer

Feature Extraction	Model	Accuracy	Precision	Recall	F1-Score
TF-IDF	**RF**	98.00%	99.00%	98.00%	98.00%
	MNB	96.11%	98.00%	92.00%	94.00%
	LR	97.00%	98.00%	98.00%	98.00%
	DT	96.00%	94.00%	97.00%	96.00%
	SVC	97.00%	98.00%	98.00%	98.00%
	XGBoost	96.11%	95.00%	97.00%	96.00%
Count Vectorizer	**RF**	97.74%	98.00%	98.00%	98.00%
	MNB	97.00%	98.00%	98.00%	98.00%
	LR	94.45%	93.00%	96.00%	95.00%
	DT	97.00%	98.00%	98.00%	98.00%
	SVC	96.11%	97.00%	97.00%	97.00%
	XGBoost	97.22%	98.00%	98.00%	98.00%

The results depict that the random forest classifier (RF) achieved the best performance among all the models with an overall test accuracy of 98% using TF-IDF and 97.74% using count vectorizer. In contrast, the lowest test accuracy of 96% is achieved by decision tree using TF-IDF and the lowest test accuracy of 94.45% is achieved by logistic regression using count vectorizer. Hence, the RF classifier was applied later to classify and predict the sentiment of the unlabelled tweets. Figure 6 shows the confusion matrix of RF model validated with a stratified 90–10% split validation scheme.

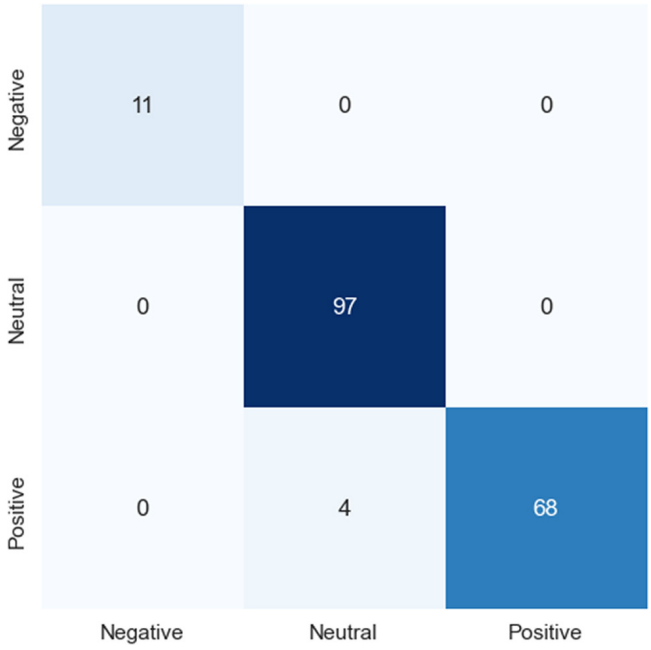

Fig. 6. Confusion matrix of the random forest classifier (RF) with TF-IDF

Figure 7 presents a word cloud of the most frequent words in the dataset. The size of a word shows how important it is. The bigger word is, the more frequently it is presented. The most frequently words shown in the figure include 'Université', 'universite, 'mohammed', 'rabat', 'science', 'cadi', 'ayyad', 'maroc', 'marrakech', 'étudiant', 'faculté', 'hassan', 'sidi', 'ben', 'abdellah', 'fès', and 'recherche'.

Fig. 7. Word cloud of frequent words in the dataset

4.3 Results of the real-time phase

The results of applying random forest model to real-time tweets related to 12 Moroccan public universities are described and discussed in this subsection. Figure 8 is a line chart that illustrates the number of tweets received per day and per sentiment about each Moroccan public university in real-time from May 25, 2022, to September 16, 2022. A large number of tweets is collected during the period between May 26, 2022, and June 10, 2022.

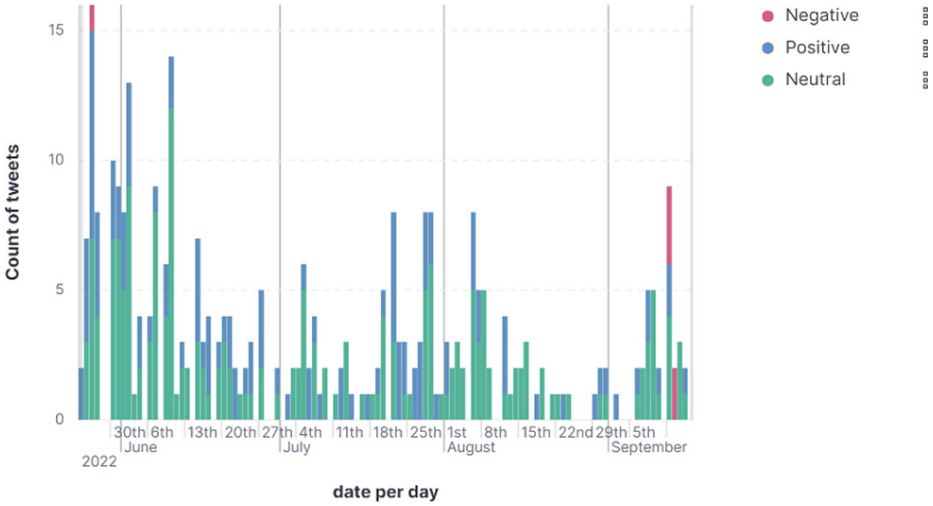

Fig. 8. Bar chart showing the number of real-time tweets received per day and per sentiment

Figure 9 is a stacked horizontal bar chart representing the sentiment of each Moroccan public university. We can see clearly from the figure that most of the real-time tweets are about Mohammed V University in Rabat, followed by Abdelmalek Essaadi University and Cadi Ayyad University. In general, the people's sentiments about Moroccan universities are positive and neutral.

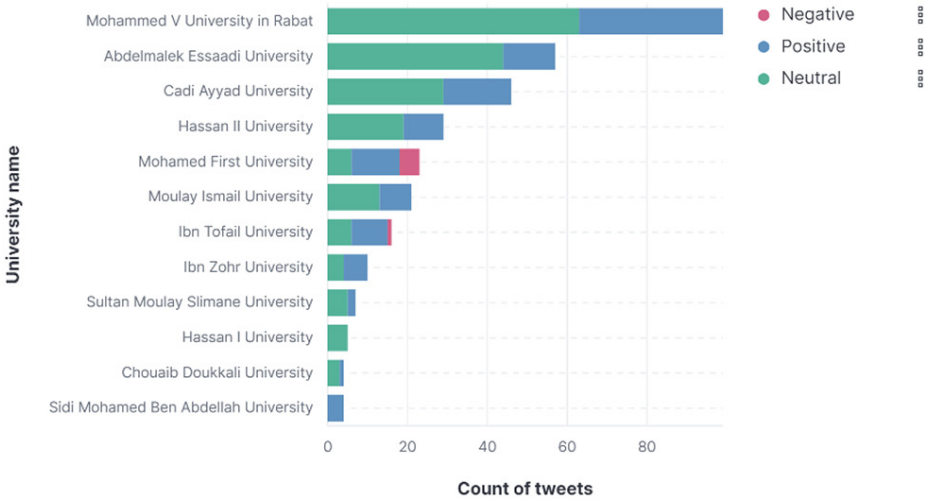

Fig. 9. Stacked horizontal bar chart showing the predicted sentiment of Moroccan universities

Figure 10 is a pie chart showing the percentage of people's sentiments towards the twelve Moroccan public universities. It is seen that 61.37% of the real-time collected tweets were neutral, 36.76% of them were positive, and 1.87% were negative. Hence, among 36.76% of the positive tweets, Mohammed V University in Rabat got the highest number of positive tweets forming 30.51%, followed by Cadi Ayyad University with 14.41%, then Abdelmalek Essaadi University with a percentage of 11.02% of positive tweets. In contrast, Chouaib Doukkali University got the lowest number of positive tweets, as shown Figure 11.

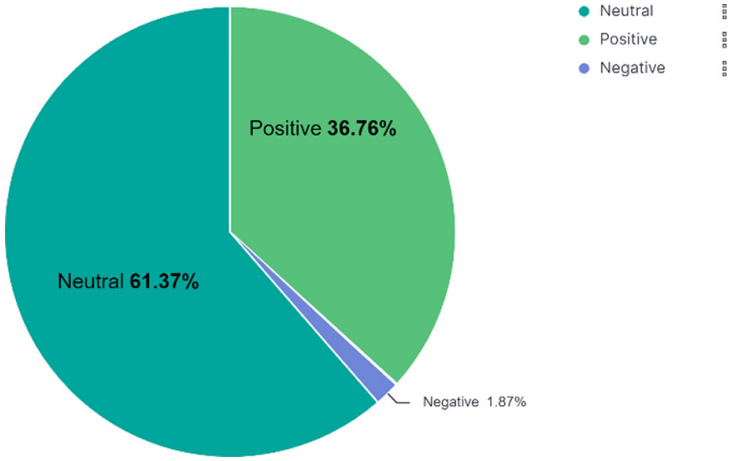

Fig. 10. Distribution of predicted sentiments of tweets related to Moroccan universities

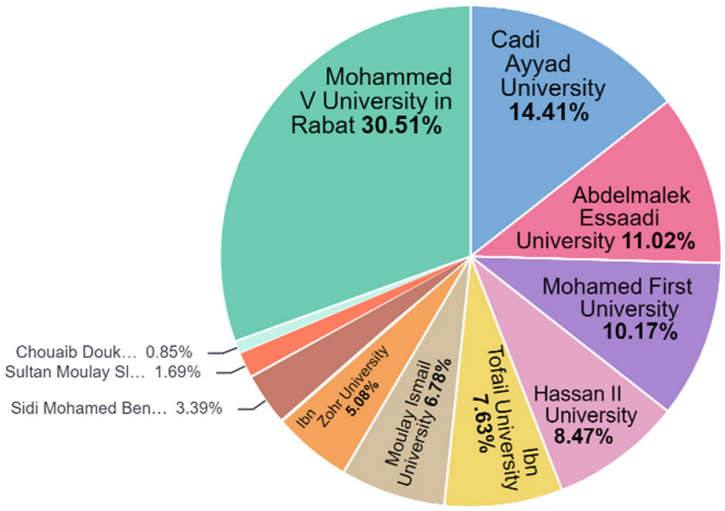

Fig. 11. Percentage of positive tweets per university

Figure 12 shows a map of geo-located tweets collected in real-time. A tweet's exact location is expressed in latitude and longitude coordinates. A close look at the map reveals that the majority of tweets concentrate on Morocco followed by Spain, France, and United States.

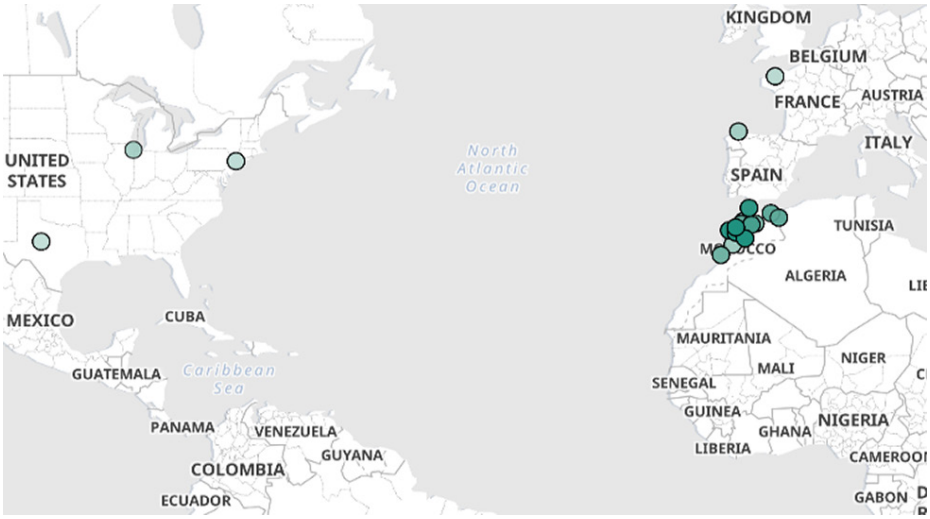

Fig. 12. Geographical distribution of the real-time tweets

Figure 13 depicts the Kibana dashboard used to track sentiments of tweets related to the twelve Moroccan public universities in real-time. It shows the total number of collected tweets and the five graphs shown above in order to real-time visualize the data.

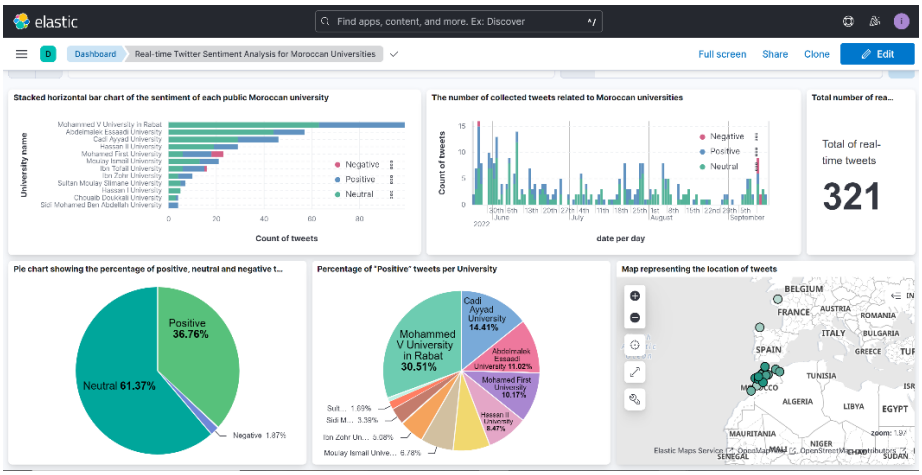

Fig. 13. Screenshot of the dashboard produced by Kibana for visualizing the real-time tweets

5 Conclusion and future work

Our study introduces a system for Twitter sentiment prediction regarding twelve Moroccan public universities in real-time. It consists of Twitter Streaming API for data collection, Apache Kafka for ingestion, Apache Spark for real-time sentiment analysis prediction, Elasticsearch, and Kibana for real-time data exploration and visualization. The proposed system is composed of two main phases: an offline sentiment analysis phase and a real-time prediction phase. We collected 1798 historical French tweets about Moroccan universities between June 11, 2009, and May 24, 2022 using Twint. The collected tweets got annotated as 'Positive', 'Negative', and 'Neutral' and then pre-processed in order to be used to train six machine learning algorithms (RF, MNB, LR, DT, SVC, and XGBoost) with TF-IDF and count vectorizer feature extraction techniques. We found that the RF classifier achieved the best test accuracy of 98% using TF-IDF. Hence, this model was used to classify 321 real-time tweets collected from May 25, 2022, to September 16, 2022. The results of the online phase reveal that 61.37% of the real-time collected tweets related to twelve Moroccan public universities were neutral, 36.76% were positive, and 1.87% were negative. Then, from 36.76% of the positive tweets, Mohammed V University in Rabat got the highest number of positive tweets, followed by Cadi Ayyad University and Abdelmalek Essaadi University. In addition, the majority of the tweets are concentrated on Morocco, followed by Spain, France, and United States. The obtained results can be used by the Ministry of higher education, scientific research and innovation of Morocco for decision making. As for future works, we intend to include other languages, such as English and Arabic.

6 References

[1] https://www.statista.com/forecasts/1146722/twitter-usersin-the-world [Accessed: May 20, 2022].

[2] Htet, H., Khaing, S. S., Myint, Y. Y. (2019). Tweets sentiment analysis for healthcare on big data processing and IoT architecture using maximum entropy classifier. Big Data Analysis and Deep Learning Applications, 744: 28–38. https://doi.org/10.1007/978-981-13-0869-7_4

[3] Machuca, C. R., Gallardo, C., Toasa, R. M. (2021). Twitter sentiment analysis on Coronavirus: machine learning approach. J. Phys. Conf. Ser., 1828 (1): 012104. https://doi.org/10.1088/1742-6596/1828/1/012104

[4] Riadsolh, A., Lasri, I., ElBelkacemi, M. (2020). Cloud-based sentiment analysis for measuring customer satisfaction in the Moroccan banking sector using Naive Bayes and Stanford NLP. J. Autom. Mob. Robot. Intell. Syst., 14(4): 64–71. https://doi.org/10.14313/JAMRIS/4-2020/47

[5] Vidya, N. A., Fanany, M. I., Budi, I. (2015). Twitter sentiment to analyze net brand reputation of mobile phone providers. In 3th Information Systems International Conference, Surabaya, Indonesia, 519–526. https://doi.org/10.1016/j.procs.2015.12.159

[6] Omar, M. F., Mahathir, N. H., Mohd Nawi, M. N., Zulhumadi, F. (2019). Prototype development and pre-commercialization strategies for mobile based property analytics. International Journal of Interactive Mobile Technologies (iJIM), 13(10): 198–204. https://doi.org/10.3991/ijim.v13i10.11309

[7] Paolanti, M. et al. (2021). Tourism destination management using sentiment analysis and geo-location information: a deep learning approach. Inf. Technol. Tour., 23(2): 241–264. https://doi.org/10.1007/s40558-021-00196-4

[8] Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., Singh, K. P. (2020). Analysis of political sentiment orientations on Twitter. In International Conference on Computational Intelligence and Data Science (ICCIDS 2019), 167:1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

[9] Ergul Aydin, Z., Kamisli Ozturk, Z., Erzurum Cicek, Z. I. (2021). Turkish sentiment analysis for open and distance education systems. Turk. Online J. Distance Educ., 124–138. https://doi.org/10.17718/tojde.961825

[10] Dina, N. Z., Yunardi, R. T., Firdaus, A. A., Juniarta, N. (2021). Measuring user satisfaction of educational service applications using text mining and multicriteria decision-making approach. International Journal of Emerging Technologies in Learning (iJET), 16(17): 76–88. https://doi.org/10.3991/ijet.v16i17.22939

[11] Ulfa, S., Bringula, R., Kurniawan, C., Fadhli, M. (2020). Student feedback on online learning by using sentiment analysis: a literature review, In 6th International Conference on Education and Technology (ICET), Malang, Indonesia, 53–58. https://doi.org/10.1109/ICET51153.2020.9276578

[12] Ouatik, F., Erritali, M., Ouatik, F., Jourhmane, M. (2021). Students' orientation using machine learning and big data. International Journal of Online and Biomedical Engineering (iJOE), 17(01): 111–119. https://doi.org/10.3991/ijoe.v17i01.18037

[13] Baragash, R., Aldowah, H. (2021). Sentiment analysis in higher education: a systematic mapping review. J. Phys. Conf. Ser., 1860(1): 012002. https://doi.org/10.1088/1742-6596/1860/1/012002

[14] Kamisli Ozturk, Z., Erzurum Cicek, Z., Ergul, Z. (2017). Sentiment analysis: an application to Anadolu university. Acta Phys. Pol. A, 132(3): 753–755. https://doi.org/10.12693/APhysPolA.132.753

[15] Zaharia, M. et al. (2016). Apache Spark: a unified engine for big data processing. Commun. ACM, 59(11): 56–65. https://doi.org/10.1145/2934664
[16] Kreps, J., Narkhede, N., Rao, J. (2011). Kafka: a distributed messaging system for log processing. In 6th International Workshop on Networking Meets Databases (NetDB 2011), 7.
[17] Gormley, C., Tong, Z. (2015). Elasticsearch: the definitive guide.
[18] Neethu, M. S., Rajasree, R. (2013). Sentiment analysis in Twitter using machine learning techniques. In 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp.1–5. https://doi.org/10.1109/ICCCNT.2013.6726818
[19] Goel, A., Gautam, J., Kumar, S. (2016). Real time sentiment analysis of tweets using Naive Bayes. In 2nd International Conference on Next Generation Computing Technologies (NGCT), 257–261. https://doi.org/10.1109/NGCT.2016.7877424
[20] McCallum, A., Nigam, K. (1998). A comparison of event models for naive bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization.
[21] Chikersal, P., Poria, S., Cambria, E. (2015). SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, 647–651. https://doi.org/10.18653/v1/S15-2108
[22] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3): 273–297. https://doi.org/10.1007/BF00994018
[23] Coletta, L. F. S., Da Silva, N. F. F., Hruschka, E. R., Hruschka, E. R. (2014). Combining classification and clustering for tweet sentiment analysis. In Brazilian Conference on Intelligent Systems, Sao Paulo, Brazil, 210–215. https://doi.org/10.1109/BRACIS.2014.46
[24] Huq, M. R., Ali, A., Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. Int. J. Adv. Comput. Sci. Appl., 8(6): 19–25. https://doi.org/10.14569/IJACSA.2017.080603
[25] Altrabsheh, N., Cocea, M., Fallahkhair, S. (2014). Learning sentiment from students feedback for real-time interventions in classrooms. In Adaptive and Intelligent Systems, Springer International Publishing, 8779: 40–49. https://doi.org/10.1007/978-3-319-11298-5_5
[26] Mujahid, M. et al. (2021). Sentiment analysis and topic modelling on tweets about online education during COVID-19. Appl. Sci., 11(18): 8438. https://doi.org/10.3390/app11188438
[27] Waheeb, S. A., Khan, N. A., Shang, X. (2022). Topic modelling and sentiment analysis of online education in the COVID-19 era using social networks based datasets. Electronics, 11(5): 715. https://doi.org/10.3390/electronics11050715
[28] Balachandran, L., Kirupananda, A. (2017). Online reviews evaluation system for higher education institution: an aspect based sentiment analysis tool. In 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2017), Malabe, Sri Lanka, 1–7. https://doi.org/10.1109/SKIMA.2017.8294118
[29] Alruily, M., Shahin, O. R. (2020). Sentiment analysis of Twitter data for Saudi universities. Int. J. Mach. Learn. Comput., 10(1): 18–24. https://doi.org/10.18178/ijmlc.2020.10.1.892
[30] Ho, T. K. (1995). Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, 278–82.
[31] AL-Rubaiee, H., Qiu, R., Alomar, K., Li, D. (2016). Sentiment analysis of Arabic tweets in e-learning. J. Comput. Sci., 12(11): 553–563. https://doi.org/10.3844/jcssp.2016.553.563
[32] Pratama, A. (2020). How to Scrape Tweets from Twitter with Python Twint. Available online: https://medium.com/analyticsvidhya/how-to-scrape-tweets-from-twitter-with-python-twint-83b4c70c5536 [Accessed: March 2, 2021].
[33] Bird, S., Klein, E., Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. OReilly Media Inc.
[34] Loria, S. (2018). Textblob Documentation. Release 0.15, 2.

[35] Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. https://doi.org/10.1108/00220410410560582
[36] Wu, X., Kumar, V., Quinlan, J. R. et al. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1): 1–37. https://doi.org/10.1007/s10115-007-0114-2
[37] Duda, R. O., Hart, P. E. (1973). Pattern classification and scene analysis. John Wiley and Sons.
[38] Cox, DR. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 120(2): 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
[39] Chen, T., Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785
[40] Hunt, P., Konar, M., Junqueira, F. P., Reed, B. (2010). ZooKeeper: wait-free coordination for internet-scale systems. 14.
[41] Pedregosa, F., Varoquaux, Gael, Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct): 2825–2830.

7 Authors

Imane Lasri is a PhD student at the Laboratory of Conception and Systems (Electronics, Signals, and Informatics), Faculty of Sciences Rabat, University Mohammed V in Rabat, Morocco. She received a Master's degree in Big Data Engineering from the Faculty of Sciences Rabat. She received the awards of excellence of the major winners from the Mohammed V University in Rabat in 2019. Her current field of research is pattern recognition applied to higher education using machine learning and deep learning algorithms (email: imane_lasri@um5.ac.ma).

Anouar Riadsolh is a Professor at the Faculty of Sciences Rabat, University Mohammed V in Rabat, Morocco. He received his PhD in Computer Science from the Faculty of Sciences Rabat. He is a member of the Laboratory of Conception and Systems (Electronics, Signals, and Informatics). His current research interests are focused on data mining, big data, and machine learning (email: a.riadsolh@um5r.ac.ma).

Mourad Elbelkacemi is a Professor at the Faculty of Sciences Rabat, University Mohammed V in Rabat, Morocco. He received his PhD in Computer Science. He is a member of the Laboratory of Conception and Systems (Electronics, Signals, and Informatics). His main research interests are focused on electronics, education, and data mining (email: mourad_prof@yahoo.fr).

Article submitted 2022-10-11. Resubmitted 2023-01-19. Final acceptance 2023-01-23. Final version published as submitted by the authors.