A NOVEL LOGISTICS TRANSPORTATION OPTIMIZATION APPROACH THROUGH THE DETECTION OF ROAD TRAFFIC EVENTS USING APACHE SPARK BIG DATA FRAMEWORK, MACHINE LEARNING AND SOCIAL BIG DATA

Mouammine Zakaria, University of Hassan the Second Casablanca
H. Khoulimi, Applied Mathematics and Computing Laboratory, Higher Normal School Casablanca
A. Ammoumou, University of Hassan the Second Casablanca
B. Nsiri, National School of Arts and Crafts of Rabat (ENSAM) Mohammed V University in Rabat

ABSTRACT

This paper proposes an innovative model to enable intelligent transportation which rationalizes logistics/transportation cost and increases subsequently the value independently to whether or not there is an intelligent infrastructure. The proposed model is an operational framework for automatic monitoring of road traffic state and transportation conditions through a near-real time detection of related events. Facebook and twitter social networks are used as a source of social data. To deal with comments/tweets written in Moroccan Dialect a set of tools related to big data architecture are utilized respectively, Machine Learning Algorithms and Apache spark framework.

We applied our model to monitor the traffic on the city of casablanca which is the larger city in morocco, the obtained results reveal that our method is able to enable intelligent transportation, more precisely it can detect automatically road traffic events, without any prior information, this has allowed the real time traffic operation monitoring, To the best of our knowledge, this is the first work addressing traffic event detection from tweets written in Moroccan dialect language using machine learning and Apache Spark big data platform.

Keywords: Big Data, Smart Logistic Transportation, Apache Spark, Machine Learning, Support Vector Machine, Naïve Bayes, Logistic Regression, Social Big Data

INTRODUCTION

Within the opening of markets into an international dimension, companies had to get adapted and reformed so as to remain competitive. Mastering supply chain is considered as a vital factor for any industrial company, logistics service provider, 3PL (third-party logistics) or any other participant. one of the principal function of SC is transportation (Carel, 2019), its mastering increases notably the added value and reinforces the SC effectiveness which impacts positively partnerships between multiple SCs (Tseng, Yue & Taylor, 2005). Within the era of new technologies as many aspects of our life has changed, industry also has been impacted and got transformed to Industry 4.0, where companies have access to a variety of advanced technologies such as Iot, Sensors, RFID, Smart devices, online social network, crowdsourcing, etc., (Barreto, 2017) Big data is one of these great technologies,

In some areas of the world logistics stakeholders are taking advantages of these new technologies in order to improve their transportation. However, smart mobility or smart logistics transportation is still a new comer aspect even for the developed countries, it is actually

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

conditioned by the implementation of the smart city environment (Nesmachnow, 2017), indeed, the huge cost of implementing an intelligent infrastructure like smart city is still challenging to afford by the most of developing countries (Galati, 2018), adding to this the need of making many upstream structural changes on multiple dimensions of the society (AlEnezi, 2018), hence the need to look for an alternative solution which ensures the effectiveness and the feasibility, such as what authors proposed in these papers (Mouammine, 2020; Mouammine) .

The accessibility to a massive amount of live data posted by millions of users on social networks. this is because of their accessibility by everyone from anywhere over the world, it's the best solution to monitor the traffic flow and it's a low coast one comparing to the IoT solution which requires an investment on the installation of expensive sensors.

As people's mentality increases for sharing traffic information, Twitter and Facebook have become the popular social networks for sharing their information. Indeed, they became a powerful data source for smart transportation (Lavanya, 2020), since then, they can be used as a sensor for flow forecasting, traffic management, and event detection. From a data mining perspective, detecting events from unstructured and rapidly changing social data is a difficult task. Therefore, the volume and variety of social data requires advanced techniques of big data architecture which are able to manage and analyze the massive data, Then, they extract useful information needed to predict the future observation about traffic and road conditions.

Moreover, the challenging aspect in this automatic event detection model is the language of comment/tweet. Already many in this domains for monitoring the road traffic using the social media by analyzing text from different language, like Chinese (Conway, 2019), Japanese (Rahman, 2019), and Italian (Polignano, 2019), As we focused in this research on tweets written in Moroccan dialect "Darija", we have encountered a serious challenge due to the diversity and complexity of this dialect compared to MSA (Modern Standard Arabic) (Tachicart, 2021).

To sum up, in the recent years several approaches were proposed to automate event detection just by analyzing social data, but until now and to the best of our knowledge there is no work addressing how to detect event transportation using Moroccan dialect language and social big data to optimize transportation.

This paper presents a model based on big data architecture for enabling smarter transportation through the analysis of gathered "comments/tweets" written in Moroccan dialect in order to detect traffic related events. we combine apache Spark framework and some algorithms of machine learning (ML), respectively, Support Vector Machine (SVM) (Preda, 2018), Naïve Bayes(NB) (Granik, 2017) and logistic regression (Aborisade, 2018) so as to filter comments/tweets into correct data and salts one. Subsequently other classifiers are utilized to detect other types of events including roadwork, accident, road closure, traffic condition and other events natural or human.

Our paper is organized as follow, in section 2 we present a review to the related work. In section 3 we explain the methodology followed to curry on our study. while the section 4 is designed to describe the proposed approach and we give in the section 5 a conclusion including some of our future perspectives.

Why ITS are Important

ITS refer to the integration of information and communication technologies in the field of management of problems encountered in conventional transport systems. They use innovative technologies so as to improve safety, effectiveness, efficiency, accessibility and sustainability of the transport network with the ensuring of an optimized transportation capacity. In addition to the significant impact of the ITS on reducing the total cost of a product, they provide more other key benefits such as:

a. Increasing people and goods safety by guiding the drivers using developed outputs such voice as well as sending an alert to the user about the traffic ahead, so that the user may act accordingly.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

b. Information sharing, ITS provide a set of messages format to send the report of the road to the users such as, Dynamic Message Signs (DMS) (Banerjee, 2019), Variable Message Signs (VMS) (Ma, 2020), and Highway Advisory Radio (HAR) (Sandt, 2017).
c. Mobility and convenience.

Related Works

Analyzing social information in Arabic language for event detection is limited compared to what is done in other languages. so in this section, we cited some existing works that deal with road traffic event detection basing on social media data.

Authors in paper (Shafiq, 2020) extracted information related to traffic using lexicon-based and rule-based techniques. to classify Thai tweets about traffic into six categories include announcement, accident, orientation, question, sentiment and request these applied machines learning classifier based on Naive Bayes Model.

In paper (Lan, 2020) author detect events related to road traffic by analyzing tweets and built a classification model to transform the tweets into traffic related and non-traffic related using methods logistic regression with stochastic gradient descent. for detecting events, they identify the most frequent terms among the traffic related tweets.

Work (McDermott, 2018) adopts the use of the training matrix to classification, this matrix contains the selected terms and their corresponding TF-IDF (Term Frequency-Inverse Document Frequency) weights. However, the model was trained on a small database that's contain about 3700 Arabic tweets to detect one type of events which is a high-risk flood.

Salas, et al., (Rettore, 2020) propose a framework for the real-time detection of traffic events from tweets in English language using Apache Spark and Python machine learning algorithms. Additionally, they used the SVM classification algorithm and classified the tweets into traffic and non-traffic related tweets.

Mathew (Soliman, 2017) proposes a framework 'logistic distribution' for generalization to LaPlace using the framework of Kozubowski & Podgorski, (2000). With the help of the said study, the author's logistic distribution is worth noticing in the present study.

Problem Definition

Providing a smart transportation basing or ITS requires an important investment on the technological ground and the associated tools, the funding of such project might not be affordable for many countries, adding to this that many change on multiple levels should take place, for instance, these are the basics factors that should be ensured first before implementing a smart transportation (Mfenjou, 2018)

- **Smart governance**: Smart governance is obtained by offering E-democracy, access and open data, transparency, encouraging public research and development and education, continuing to increase the investment on roads and infrastructure, reinforcing social and relational capital so as all social classes should benefit from the success of the high-technology.
- **Smart people**: The higher education levels lead to a better environment for new enterprises, creating new knowledge, higher qualified workers, jobs and business opportunities.
- **Smart environment**: Smart environment is efficiency and sustainability, ecologically sustainable enterprises, renewable energy production, urban tooling and pollution control. It is related to the economy on loop.
- **Smart living**: Smart living means the best conditions of life (security and quality of life), meaning living with healthy people and healthy buildings in the best conditions. So, city utility infrastructure (water/electricity/heating network, lighting, waste disposal…) and smart sensor network (internet for everything project).

Moreover, the funding is still the most challenging part of the implementing of connected ITS infrastructure, we highlight the example of morocco which wanted to implement

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

a smart city project , thus, due to the colossal cost of the project, the letter has to agree with many actors to ensure the financing of Tangier Tech project which was estimated to cost $10 billion (TheArabWeakly, 2021).

To overcome this constraint, indeed, we aim through this work to propose an alternative solution to the technological ground required for enable intelligent transportation, only relying on low-cost data sources and big data architecture.

Data Sources

The data sources have been used for this model for feeding the proposed framework, are offered by the two widely used social networks (Table 1), twitter and Facebook, through their APIs.

Twitter API

The Twitter APIs give users access to four key items that were considered for this project. These are some of them:

- **Tweets**: represent the most basic entity and can be embedded, liked, disapproved, and removed. coordinates, creation time, tweet id, language, place, and content are the important tweet fields streamed/retrieved for analysis.
- **Users**: They can tweet, follow, make lists, have a personal timeline, or be mentioned, however, Account creation timestamp, description, number of followers, number of friends, geotagging, account id, language, location, are all the important user fields to evaluate on this study.
- **By giving** a place id when tweeting, specific, named locations can be associated to Tweets. Place-related tweets are not necessarily sent from that area, but they could be about it. Places may be found and Tweets can be obtained based on their id. indeed, The place elements are considered attributes.

Facebook Graph API

The Graph API is the best way to insert and retrieve data in the Facebook platform. It is an HTTP-based API that allows apps to use programming to query data, publish news, manage ads, import photos and perform a wide range of other tasks. The name of the API Graph is inspired by the idea of a "social graph": A representation of information on Facebook. This graph is composed of the following elements:

- **Nodes**: represent essentially individual objects, such as user, photo, page or a comment.
- **Edges**: connections between a collection of objects and a single object, such as photos on a page or comments on a photo.
- **Fields**: data like an object, such as a user's birthday or the name of a page.

Table 1
VOLUME AND VELOCITY FOR THE AVAILABLE DATA SOURCES [26] [27].

Data source	Volume	Velocity	Variety
Twitter	500 million tweets/day	1-5 minute	json format
Facebook	734 million comments/day	1 minute	json format

METHODOLOGIE

We present on the fig 1, our proposed Multi layers architecture of the novel approach of traffic event detection based on social big data, Apache Spark, machine learning, this architecture contains six main layers: Data collection and storage layer, Data pre-processing layer, Feature extractor layer, Tweet/comment filtering layer, Event detection layer, and results visualization.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

Firstly, once data are collected on JSON format, they get stored as objects (Kumar, 2017) into a MongoDB database (MongoDb, 2021) , we proceed to their filtering after removing duplicates, we categorize comment/tweet to labeled and unlabeled dataset by setting 1 for relevant and 0 for irrelevant. Then we remove noise and prepare the data for classification on the pre-processing steps, we got as a result a list of normalized and cleaned tokens on which we apply TF-IDF techniques (Kim, 2019), Bug of words (BoW) serves also for the same objective (Pimpalkar, 2020), however we preferred to use TF-IDF model for its reliability and performance. The labeled comment/tweets are used to set up and train a classifier to filter the later into relevant or irrelevant.

For event detection, we relied on three models, using three different supervised classification algorithms, then we apply the four widely known evaluation metrics, which are, precision, accuracy, recall, and F-score in order to evaluate the adopted models and then to select the best algorithms among them, after that we adopt the trained model that allows achieving the higher performance. In the following step, a subset of relevant tweets is manually labeled and utilized to train and create other event classifiers. The trained classifiers are evaluated and then used to detect event. Finally, we visualize the results and make sure of the reliability of the adopted classifier, by searching in the official sources such as newspaper platforms.

Furthermore, to manage the massive volume of unstructured data in the Facebook/Twitter network utilized for event detection, we rely on the Apache Spark platform, which is a distributed in-memory computing platform. We also employ the Python Machine Learning (Spark ML) package, which provides high-level machine learning APIs based on Spark DataFrame, it is based on the RDD (Resilient Distributed Datasets) (Jonnalagadda, 2016) which allow for higher performance and higher speed of computing, the main steps of our system are described below:

Data Gathering

This step was the most difficult party because it was needed to look for facebook/twitter pages which are dynamic in term of sharing information about event on cities, Data (tweets/comments) are collected respectively, through Twitter API and Facebook Graph API, we determine the geolocalisation to obtain the social data posted in morocco, and then we get the comments/tweets in hashtags that usually described events in cities such as '# الآن البيضاء" الدار meaning '# Casablanca now','# الآن الرباط' (‘#Rabat now’),' # أحداث طنجة' (‘#Tanger events’), we gathered all Moroccan dialect social data between july 23 and septembre 1, 2020, We opted for NoSQL databases over relational databases because our data required scalable and flexible schema-based storage. The comments/tweets are kept in MongoDB, a document-oriented database designed for storing and managing large collections of documents, MongoDB database is updated using the JSON objects obtained from our two sources. collected object have different attributes, including ‘created at,' which shows the time the tweet was posted, and the 'full text,' which provides the message content. After that, we checked for duplicate comments/tweets and erased them.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

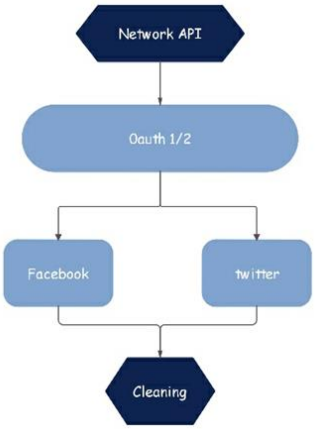


FIGURE 1
SOCIAL DATA COLLECTION MODEL

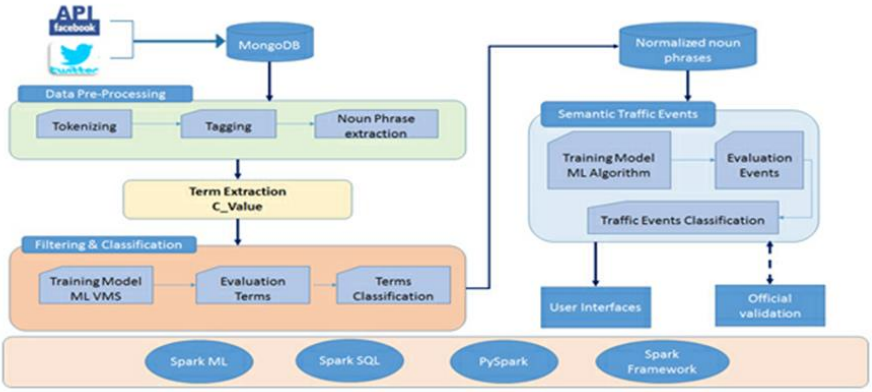


FIGURE 2
MULTI LAYERS ARCHITECTURE OF THE NOVEL APPROACH OF TRAFFIC EVENT DETECTION BASED ON TWITTER DATA, APACHE SPARK, MACHINE LEARNING

Data Pre-Processing

Arabic has complex morphology and many dialects. its complexity increases when considering the informal nature of social media text and the distinction between Modern Standard Arabic (MSA) and Dialectical Arabic (DA) such as Moroccan dialect, this is the reason why Arabic NLP has not been developed as well as English NLP (Ali, 2016), also relaying on Arabic-to-English translations give poor result. therefore, to process Moroccan dialect text, pre-processing stage is essential to reduce the volume of noise before classification because moving directly to analysis might give unreliable results. Moreover, Moroccan Arabic (darija) is sometimes written in Latin letters and sometimes in Arabic, sometimes Moroccan write in MSA, in this work we only use text written in Arabic letters, after filtering out non-Arabic content, we remove all Moroccan diacritic, the emphasis symbol which is a diacritic shaped like a small written "w" among letter, after that, the text is divided into words (tokens). The tokens are normalized to replace letter that has different forms into the basic shape. For instance, the letter (' ا ') pronounced Alif had three forms ('إ' ,'آ' ('أ', and normalized to bare Alif ('ا'), this is done with all Arab letters written in different forms. We have to drop the stop words included on gathered texts by using the Stop-Words list provided through Toolkit (NLTK) (Loper, 2002). Then, we alter the list to complete with the missing words in order to normalize the words.

Further, before starting the process of filtering and extraction we check the result of the pre-processing stage, whenever the number of tokens is equal to zero, the comment/tweet are dropped from the processing, As The English translation gives undoubtedly ineffective analysis,

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

in this work we utilize, additionally to Apache Spark OpenNLP , AraVec (Soliman, 2017) library, a word embedding open source project which aims to provide the Arabic NLP researchers with free and powerful word embedding models, this library first pre-processing step that was carried out on the collected data was filtering out non-Arabic content

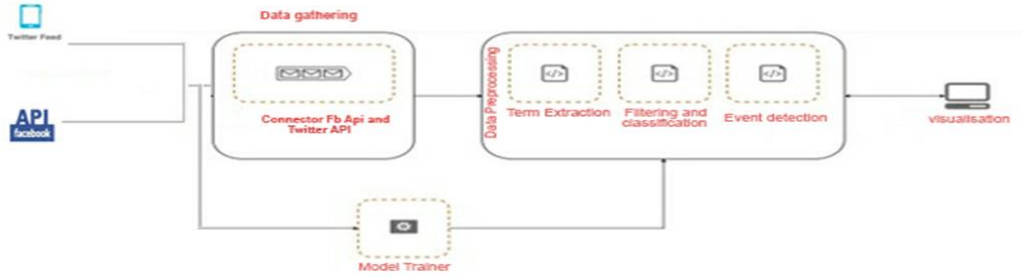


FIGURE 3
SIMPLE PIPELINE DESIGN FOR NLP PROCEDURE TO DETECT RELATED TRANSPORTATION EVENTS

Term Extraction and Selection

For term extraction we are based on the spark machine learning algorithms, we use the Term Frequency/Inverse Document Frequency TF/IDF to make this specific stage. The main objective of this technic is to find the importance of a word inside a document (comments/tweets). The function $T_f(tr, doc)T_f(tr, doc)$ is the recurrence of the appearance of term tr in a document doc, the $I_{DF}(tr, doc)I_{DF}(tr, doc)$ is measured by calculating the number of occurrence the term or its synonyms appears in the document. to calculate the IDF value we use the following equation:

Eq1 $$I_{DF}(tr, doc) = log\frac{|DOC|+1}{T_f(tr,doc)+1}\ I_{DF}(tr, doc) = log\frac{|DOC|+1}{T_f(tr,doc)+1}$$

The parameter |DOC| is the number of comments/tweets inside the collection DOC. Document Frequency $\mathrm{TF}(tr, doc)\mathrm{TF}(tr, doc)$ represents the number of the document in which the term tr is present.

Eq2 $$T_f I_{DF}(tr, doc) = T_f(tr, doc).I_{DF}(tr, doc)T_f I_{DF}(tr, doc) = T_f(tr, doc).I_{DF}(tr, doc)$$

After identifying the IDF elements, the focus is implementing the Term frequency vector using the value using CountVectorizer algorithm which obtains at first the list of 'tokens' column as input, and then it creates the vectors of token counts, the resultant TF vectors are transferred to the IDF function, the feature vectors are rescaled using IDF mode, the obtained result is stored in a column named "Features" which constitutes the input for the filtering and classification stage.

Filtering and Classification of Terms

The collected tweets are related to various topics, so we need to filter data that are related only to traffic topic, for this, we need to use a specific filter algorithm that combines spark machine learning package. We divided the manually labeled data into training (80%) and testing (the remaining 20%). After that, we use the Naïve Bayes, SVM, and logistic regression (LR) methods to develop and train the model. On the training set, the models are trained. We analyze the algorithms on the testing set to select the best one. To assess precision, accuracy, recall, and F-score, popular statistical metrics are utilized the trained classifier, we refer to traffic-related tweets as positive class and none-related tweets as negative class to clarify the meaning of these measures. these metrics employ the following four classes: True Positive (TP) refers to positive tweets that were accurately projected as positive, True Negative (TN) to

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

negative tweets that were correctly forecasted as negative, and False Positive (FP) to tweets that were incorrectly predicted as positive, and (iv) False Negative (FN) for tweets with a positive label but a negative prediction. The equations of ach matric are provided below. Eq. (3) calculates the accuracy, Eq. (4) calculates the precision (positive predictive value), Eq. (5) calculates the recall (true positive rate), and Eq. (6) calculates the F-Score.

(Eq 3) $\text{acc} = \frac{TP+TN}{TP+FP+FN+TN}$

(Eq 4) $\text{PPV} = \frac{TP}{TP+FP}$

(Eq 5) $\text{TRP} = \frac{TP}{TP+FN}$

(Eq 6) $\text{F}(\beta) = (1+\beta 2.\left(\frac{PPV.TRP}{\beta 2.PPV+TPR}\right)$

The representation of the documents as a vector allows us to see sub-areas, each one represents a specific topic that is related to our main topic "traffic topic". In the Word to Vector model, we will take a dataset and will train our model to identify the sentences or words related to our topic by comparing our term class (topic) to the predefined topics inside the vector space. However, we need to evaluate the result provided by this training model using a set of metrics such as accuracy, precision, rating score and Recall. Those metrics allow us to check if the result provided by this classifier are true and related to the traffic topic. So two classes will be identified namely positive terms (PT) and negative terms (NT).

Event Detection

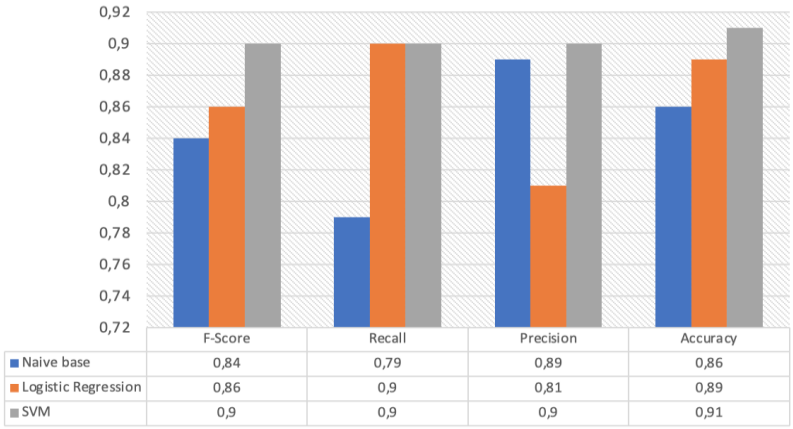
In this stage we build a classifier that we train using Naïve bayes, logistic regression and SVM. The events classifier is trained through labeling manually part of the filtered data from the previous step into eight event categories, which are Weather, Fire, Social Events, Traffic Condition, Roadwork, Road Damage, Accident, and Road Closures, these categories have positive and negative comments/tweets about the traffic condition. for some of these categories we have to consider all kind of data (positive or negative), whereas for some of them we consider only data which can affect negatively the road conditions. We've noticed that some event categories receive a lot more tweets than others. Based on the quantity of tweets, we categorized them into small-scale and large-scale events. Traffic Conditions, Roadwork, Road Damage, Accidents, and Road Closures are examples of small-scale occurrences. In comparison to Fire, Social Events, and Weather, these events receive a tiny number of tweets. As a result, we regard them as large-scale events.

Furthermore, we extract information about each event including location information using the top frequent terms since people usually refer to the event place using the hashtag. For model evaluation, we use the same evolution method explained in the stage C. in order to validate the effectiveness of our event detection approach, we extract the top vocabularies from the tweets of each detected events. Then, we use these vocabularies to search in the official news/ newspapers websites to confirm the occurrence of the events. After that, we compare the extracted information by our method including time and location with the real information in the official platforms.

RESULT AND DISCUSSION

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

In this section we present an implementation of a case study on filtering and classifying the different events related to transportation in Casablanca city through the analysis of social big data (Facebook/twitter). we compare the performance between the three used algorithms SVM, NB, and Logistic Regression algorithms used for data filtering.



	F-Score	Recall	Precision	Accuracy
Naive base	0,84	0,79	0,89	0,86
Logistic Regression	0,86	0,9	0,81	0,89
SVM	0,9	0,9	0,9	0,91

FIGURE 4
RESULT ANALYSIS FOR COMMENTS/TWEETS FILTERING

According to the fig 4, the SVM model is better than Logistic Regression and NB in term of F-score, accuracy, recall and precision. However, Logistic Regression and SVM achieved recall of 90%.

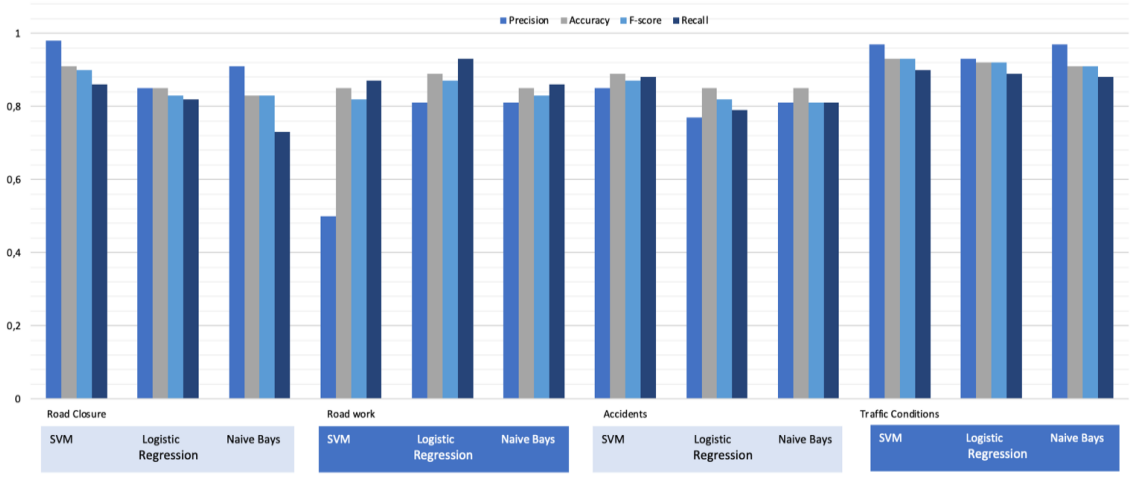
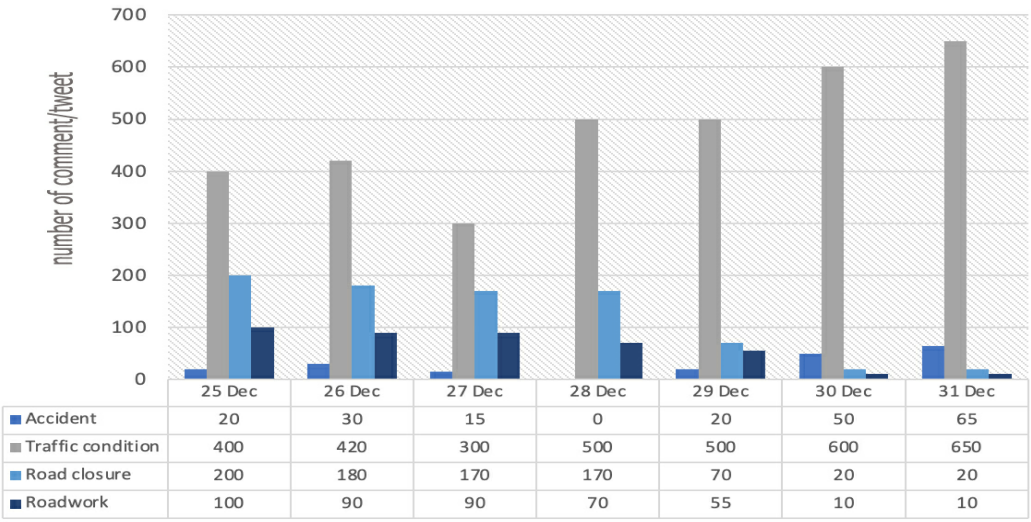


FIGURE 5
EVALUATION RESULTS FOR EVENTS CLASSIFICATION ACCURACY, PRECISION, RECALL AND F-SCORE

As mentioned in figure 5, a simulation is created to choose the best algorithm basing on the specific metrics which are Recall, Precision, Final score and Accuracy. The three algorithms show better results for each metric and give higher results. To extract the effectiveness of the three methods, we evaluate the received comments/tweets for each type of event. In our case, the SVM takes advantage of the process of computing to give better results for the events like traffic condition and accident. For the rest of the events such as road damage and road work, the LRA give us higher results. we noticed that the two technics SVM and LRA are able to achieve better results for the given events.

Furthermore, we noticed that the number of events is not fair, this one is related to the comments/tweets of users, so our dataset will receive all kinds of events and then will extract useful events according to our goals.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.



	25 Dec	26 Dec	27 Dec	28 Dec	29 Dec	30 Dec	31 Dec
Accident	20	30	15	0	20	50	65
Traffic condition	400	420	300	500	500	600	650
Road closure	200	180	170	170	70	20	20
Roadwork	100	90	90	70	55	10	10

FIGURE 6
THE NUMBER OF DETECTED SMALL SCALE EVENTS

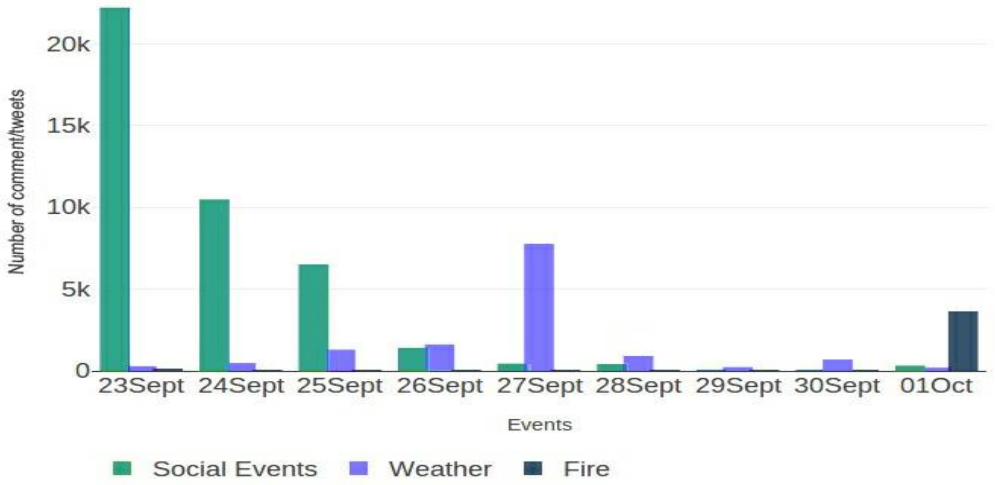
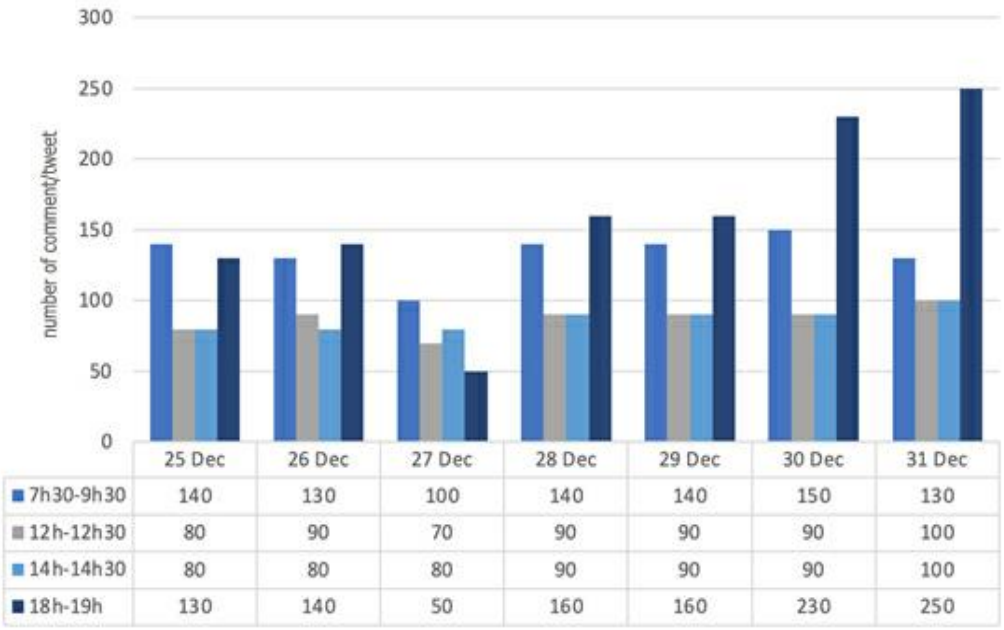


FIGURE 7
THE NUMBER OF DETECTED LARGE SCALE EVENTS.

We fetch the collected dataset and we divide events into two categories basing on the number of events and their types, we get small scales events like road closure, road work, traffic condition and accident figure 6. For large scale, we have road damage, accident and traffic condition fig 7. To validate the reliability of each event we take into consideration the number of comments/tweets that contain these terms within the use of the event location and the date of the publication, to get the number of events per day.

To detect accident events, we extract and capture events or vocabulary which include "كسيدة في الطريق بين شارع 33 و 55" Traffic accident between road 33 and 55) or كسيدة في شارع الجيش الملكي" ' (Traffic accident on the Royal Army Street) indicates that we have an accident in Royal Army Street at a specific time and date. fig 8 shows that the number of comment/ tweet telling about accident increases at the working time between 7h30 to 9h:30, 12h to 12h30, 14h to 14h30 and 18h to 19h.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.



	25 Dec	26 Dec	27 Dec	28 Dec	29 Dec	30 Dec	31 Dec
7h30-9h30	140	130	100	140	140	150	130
12h-12h30	80	90	70	90	90	90	100
14h-14h30	80	80	80	90	90	90	100
18h-19h	130	140	50	160	160	230	250

FIGURE 8
THE NUMBER OF COMMENT/TWEET DURING DIFFERENT TIME INTERVAL ON A DAY PER TYPE OF EVENT

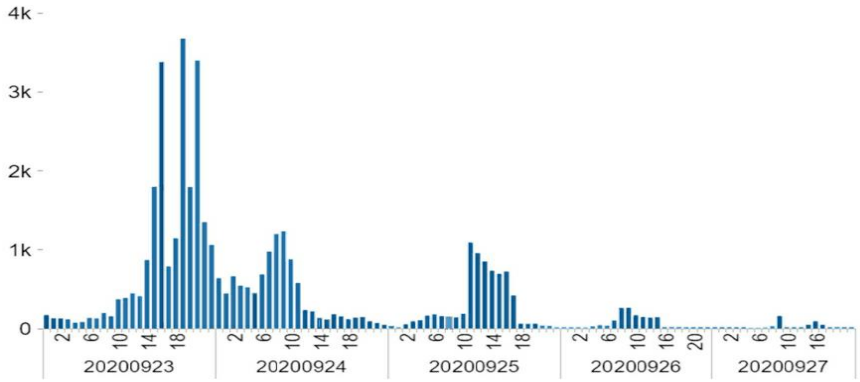


FIGURE 9
THE NUMBER OF COMMENT/TWEET PER HOUR FOR THE TOP 'SOCIAL EVENT'

For traffic condition events, we listed the top extracted jargon about traffic which are written like this: "طريق عامرة فشارع القدس سيدي معروف" (The road is full on "qouds" Street, "Sidi Maarouf"). To validate the reliability of this events, we check out on the newspaper and count the number of terms that are related to that event. As shown in Figure 8 the number of tweets related to traffic road condition in a specific period.

Fig 9 displays the number of comments/tweets related to the Social event per hour. The following terms are the top vocabulary we found written the most on the two social networks, the social Event on the 18th of novembre, (وطنى, national), (يوم, Day), (احتفالات, celebrations). The vocabulary show that the detected event is the morrocan national independency day, which includes several activities all over the country. The

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

presented results confirm the effectiveness of our approach for filtering and detecting traffic events, and by consequences optimize transportation, based only on social data.

CONCLUSION

Regarding the higher velocity of data provided by social network, road traffic event detection is still challenging for the transportation application. Hence, this paper focused on developing a novel approach to enable Road Traffic Event Detection and provide a smarter transportation.

In the followed case study, the traffic event detection has been automated using machine learning algorithms based on NLP field and Apache Spark big data framework.

Processing written Moroccan language was challenging due to the fact that it includes many languages and many ways of writing. Therefore, As the text can't serve for a direct input for classification a set of steps are required to prepare the text, select useful terms, normalize the text, and apply the TF-IDF to obtain a vector of words.

Moreover, a classifier has been trained to filter data to whether relevant or irrelevant to a transportation topic, we used the following three algorithms of machine learning, Naïve Bayes, SVM, and LR, then, we trained the other classifiers to detect the occurrence of different events related to transportation in Casablanca city.

Afterward, all transportation related events are classified, we obtain for each event the information about the category of the vent, its location and its timing , we confirm this i information by checking out on the official sources such as official websites or electronic newspaper. Experimentally, the result of the proposed method shows its ability to detect real transportation related events in real time.

Future Work

A set of improvements will be applied in our next work, on a technical aspect we will use the broker "Kafka" which will allow our model not only to register to different streaming sources of data, but also to ensure its sustainability and to enable fault tolerance. the crowdsourcing will take part as a producer of data for Kafka broker.

REFERENCES

"*Artificial intelligence and natural language processing: the Arabic corpora in online translation software," 3*(9), 59-66.

Aborisade, O., & Anwar, M. (2018). "Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 269-276.

AlEnezi, A., AlMeraj, Z., & Manuel, P. (2018). "Challenges of IoT based smart-government development," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 1-6: IEEE.

Banerjee, S., Jeihani, M., Khadem, and D.D.J.E.t.r.r. (2019). *Brown, "Units of information on dynamic message signs: a speed pattern analysis," 11*(1), 1-9.

Barreto, L., Amaral, A., & Pereira, T.J.P.M. (2017). "*Industry 4.0 implications in logistics: an overview*," *13*, 1245-1252.

Carel, L. (2019). "*Big data analysis in the field of transportation*," Université Paris-Saclay.

Conway, M., Hu, M., & Chapman, W.W.J.Y.O.M.I. (2019). "*Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data," 28*(1), 208.

Galati, S.R. (2018). "*Funding a Smart City: from concept to actuality*," in *Smart Cities*: Springer, 17-39.

Granik, M., & Mesyura, V. (2017). "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 900-903.

Jonnalagadda, V.S., Srikanth, P., Thumati, K., & Nallamala, S.H. (2016). "*A review study of apache spark in big data processing," 4*(3), 93-98.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.

Kim, S.W., Gil, J.M.J.H.C.C. (2019). "*Research paper classification systems based on TF-IDF and LDA schemes," 9*(1), 1-21.

Kumar, J., & Garg, V. (2017). "Security analysis of unstructured data in NOSQL MongoDB database," in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 300-305.

Lan, S., Tseng, M.L., Yang, C., & Huisingh, D.J.S.O.T.T.E. (2020). "Trends in sustainable logistics in major cities in China," *712*, 136381.

Lavanya, B.M.A.K. (2020). "*Social Media Data Analysis for Intelligent Transportation Systems," presented at the International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*.

Loper, E. (2002). S. J. a. p. c. Bird, "Nltk: The natural language toolkit,".

Ma, Z., Luo, M., Chien, S.I.J., Hu, D., & Zhao, X.J.P.O. (2020). "*Analyzing drivers' perceived service quality of variable message signs (VMS)*," *15*(10), e0239394.

Marr, B. (2021). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*.

Mathew, T. (2019). "*A Generalization to LaPlace and Logistics Distributions," 8*(1), 3512-3542.

McDermott, C.D., Haynes, W., & Petrovksi, A.V.J.I.J.C.S.A. (2018). "*Threat Detection and Analysis in the Internet of Things using Deep Packet Inspection*," *3*(1), 61-83.

Mfenjou, M.L., Ari, A.A.A., Abdou, W., & Spies, F.J.S.C.I., Systems, "*Methodology and trends for an intelligent transport system in developing countries*," *19*, 96-111.

MongoDb. (2021). *MongoDb website*.

Mouammine, Z., Ammoumou, A., Nsiri, B., & Bourekkadi, "*Innovative architecture based on big data and genetic algorithm for transport logistics optimization*," *98*(17).

Mouammine, Z., Ammoumou, A., Saad Ennima, E., & Nsiri, B. (n.d.). "*Big Data with Distributed Architecture Using Genetic Algorithm in Intelligent Transport Systems*."

Nesmachnow, S., Baña, S., & Massobrio, R.J.E.E.T.O.S.C. (2017). "A distributed platform for big data analysis in smart cities: combining intelligent transportation systems and socioeconomic data for Montevideo, *Uruguay," 2*(5).

Pimpalkar, A.P., & Raj, R.J.R.J.A.A.I.D.C. (2020). "*Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features*," *9*(2), 49-68.

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., & Basile, V. (2019). "Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets," in *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, 2481, 1-6.

Preda, S., Oprea, S.V., & Bâra, A.J.S. (2018). "*PV forecasting using support vector machine learning in a big data analytics context*," *10*(12), 748.

Rahman, A., & Hossen, M.S. (2019). "Sentiment analysis on movie review data using machine learning approach," in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 1-4.

Rettore, P.H., Santos, B.P., Lopes, R.R.F., Maia, G., Villas, L.A., & Loureiro, A.A.J.I.T.o.I.T.S. (2020). "*Road data enrichment framework based on heterogeneous data fusion for its," 21*(4), 1751-1766.

Sandt, A., Al-Deek, H., Rogers Jr, J.H., & Kayes, M.I.J.T.R.R. (2017). "*Using agency surveys and benefit–cost analysis to evaluate highway advisory radio as regional traveler information and communication tool*," *2616*(1), 81-90.

Shafiq, M., Tian, Z., Bashir, A.K., Jolfaei, A., & Yu, X.J.S.C. (2020). Society, "Data mining and machine learning methods for sustainable smart cities traffic classification: a survey," *60*, 102177.

Soliman, A.B., Eissa, K. (2017). S. R. J. P. C. S. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," *117*, 256-265.

Statistica. (2016). *Distribution of tweets per user per day from the Middle East and North Africa in March 2016, by country (2017 ed.)*.

Tachicart, R., Bouzoubaa, K.J.J.O.K.S.U.C. (2021). "*Moroccan Arabic vocabulary generation using a rule-based approach*,".

TheArabWeakly. (2017). *Morocco, China agree on financing for $10 billion tech city (99 ed.)*.

Tseng, Y.Y., Yue, W.L., & Taylor, M.A. (2005). "*The role of transportation in logistics chain*," 2005: Eastern Asia Society for Transportation Studies.

Citation Information: Zakaria, M., Khoulimi, H., Ammoumou, A., & Nsiri, B. (2021). A novel logistics transportation optimization approach through the detection of road traffic events using apache spark big data framework, machine learning and social big data. *Journal of Management Information and Decision Sciences, 24*(S6), 1-13.