Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BIG DATA AND MACHINE LEARNING APPROACH FOR AN EFFICIENT INTELLIGENT LOGISTICS TRANSPORTATION

[1]Z. MOUAMMINE, [1]H. KHOULIMI,[2]O. EL IMRANI, [3]M. CHRAYAH,[4]A. AMMOUMOU,[5]B. NSIRI

[1]Laboratory Industrial Engineering, Information Processing and Logistics (GITIL)

Faculty of Science Ain Chock. University Hassan II, Casablanca, Morocco

[2]FSJESTE, Abdelmalek Essaadi University, Tetouan, Morocco

[3]Ensate, Abdelmalek Essaadi University, Tetouan, 93000, Morocco

[4]Applied Mathematics and Computing Laboratory, High School of Technology University Hassan II, Casablanca, Morocco.

[5]Research Center STIS, M2CS, National School of Arts and Crafts of Rabat (ENSAM), Mohammed V University in Rabat, Morocco

E-mail:[1]zakaria.moammine@gmail.com.

ABSTRACT

Logistics is the pillar of any industrial activity, transportation and itineraries management are the essential functions of logistics, within the appearance of smart city infrastructure, intelligent transportation appeared as a cutting-edge technological newcomer, though, all researches curried out about intelligent transport/logistics require basically the existence of intelligent ground such smart city infrastructure, sensors or "IoT" so as to work, however, this is still challenging for developing countries. Thus, there is a need for an alternative framework allowing "ITS" to work independently to the infrastructure. To fulfill that demand, authors suggest in this paper an innovative model to enable smarter transportation no matter there is an intelligent ground or not. It enables automatic monitoring of road traffic state and transportation conditions via near-real time detection of road events. Facebook and twitter are used as sources of social big data that are needed for our framework. "PNL" is applied to process Moroccan Dialect texts. The obtained results prove the effectiveness of our approach, to the best of our knowledge; this is the first work treating traffic event detection from tweets written in Moroccan dialect language using machine learning and Apache Spark big data platform.

Keywords: *Big Data, Logistic transportation, Apache Spark, Smart transportation, Machine learning, Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression, Social big data*

1. INTRODUCTION

The opening of markets into an international dimension pushed companies to make changes so as to remain competitive. controlling supply chains is one of the essential functions of SC for any industrial company, logistics service provider or any stakeholder transportation [1], its mastering increases considerably the value and reinforces the SC effectiveness which affects positively the partnerships quality between multiple SCs[2]. In this era of new technologies many aspects of industrial sector have changed, the emergence of Industry 4.0 version allows companies to access to a variety of advanced technologies such as "IoT", Sensors, RFID, Smart devices, online social network, crowd sourcing, etc.,[3]They generate continuously a huge amount of data. This context mad Big data appeared as a powerful architecture allowing the use of these data.

Some logistics stakeholders take advantages of these technologies to improve their transportation services. yet, smart mobility or intelligent logistics transportation is a new comer aspect even for the developed countries, it actually requires many bases related to the implementation of the smart city environment[4],indeed, the huge cost of implementing an intelligent infrastructure like smart city remains difficult to afford by the majority of developing countries[5], adding to this the need to

do many upstream structural changes on multiple aspects of the society[6].searching for an alternative solution which would ensure the effectiveness and the feasibility was the motivation behind what the authors proposed in these papers[7][8]

The use of the massive amount of live data posted by millions of users on social networks has become possible thinks of their opening to everyone from anywhere all over the world, it's the optimal solution to keep an eye on the traffic flow, it's cheap compared to the IoT option which require investment for the establishing 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[9], since then, they were 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 Chines [10], Japanese [11], and Italian [12], 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)[13].

To sum up, numerous models have been offered in recent years to automate event detection by analyzing social data, but to our knowledge, no work has been done on how to identify event transportation utilizing Moroccan dialect language and social big data to optimize transportation.

This paper presents a new approach based on big data architecture for enabling smarter transportation via the analysis of collected "comments/tweets" written in Moroccan dialect for detecting traffic related events. we combine apache Spark framework and some algorithms of machine learning (ML), for instance , Support Vector Machine (SVM), Naïve Bayes(NB)[14] and logistic regression[15] so as to filter comments/tweets into correct data and salts one. Subsequently other classifiers are utilized to detect other types of events including road work, accident, road closure, traffic condition and other natural or human events.

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.

2. 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.

the authors of paper [16] used lexicon-based and rule-based approaches to extract traffic-related data. These employed machines learning classifier based on Naive Bayes Model to categorize Thai tweets concerning traffic into six categories: announcement, accident, orientation, inquiry, sentiment, and request.

In paper [17]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 [18] 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. [19] 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.

3. WHY ITS ARE IMPORTANT

The term "ITS" refers to the use of information and communication technology in the management of issues that arise in traditional transportation systems. They employ cutting-edge technology to increase the transportation network's safety, efficacy, efficiency, accessibility, and long-term viability while assuring optimal transit capacity. ITS have a huge impact on decreasing the entire cost of a product, but they also bring additional crucial benefits such as:

- ✓ Increasing people and goods safety: through guiding drivers by 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.
- ✓ Information sharing: they provide a set of format types of messages to send the report of the road to the users such as, Dynamic Message Signs (DMS)[20], Variable Message Signs (VMS)[21], and Highway Advisory Radio (HAR)[22].
- ✓ Mobility and convenience.

4. PROBLEM DEFINITION

Establishing ITS requires a big changes in term of the technological ground as well as the associated tools, the funding of such project might not be affordable for some countries, indeed, these are the basics items that should be ensured first before implementing a smart transportation[23].

- ✓ Smart governance: Smart governance is obtained by offering E-democracy, access and open data, transparency, encouraging public research and development and education,
- ✓ Smart people: Higher education levels lead to a better environment for new companies, producing knowledge and enhancing the way of living, higher qualified workers, jobs and business opportunities.
- ✓ Smart environment: Smart environment, called also to the economy on loop is based on efficiency and sustainability, ecologically sustainable enterprises, renewable energy production, urban tooling and pollution control.
- ✓ Smart living: this means providing the best conditions for living with healthy people and healthy buildings in the best conditions, city utility infrastructure (water/energy/heating network, lighting, waste equipment…) and smart sensor network.

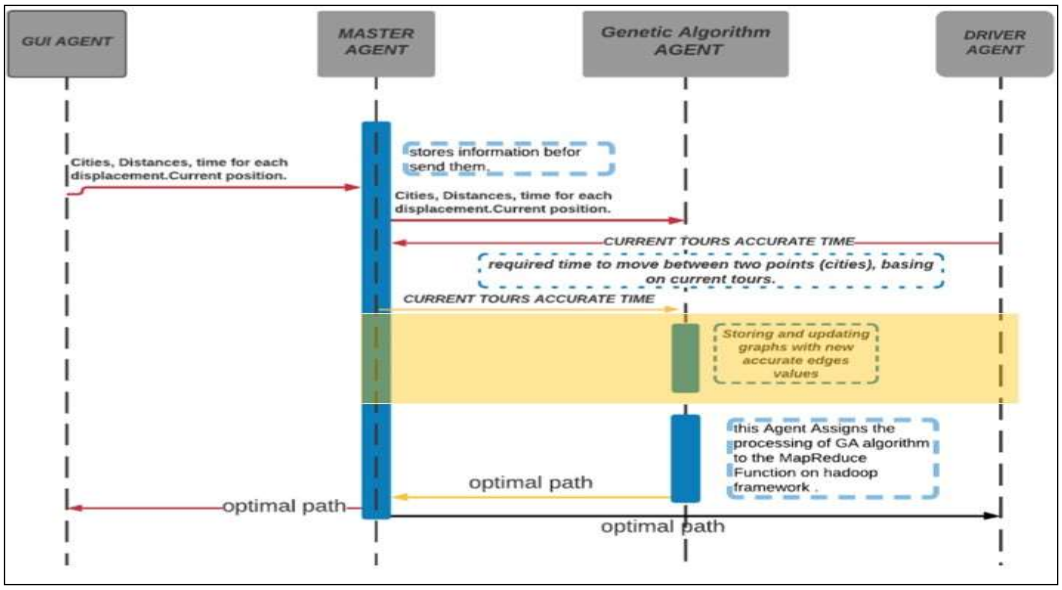

Figure1: Interaction Diagram for Distributed Model for Solving Traveling Salesman Problem [7].

Furthermore, the funding is still the most challenging part of the implementing of connected infrastructure, we highlight the example of morocco which wanted to implement a smart city project, thus, due to the colossal cost of the project, the letter had to cooperate with many actors to ensure the financing of Tangier Tech project which was estimated to cost $10 billion[24].

We overcome this constraint by proposing an alternate solution to the technological groundwork required for intelligent mobility implementation, based solely on low-cost data sources and big data architecture.

This work is complementary to a multi-agent framework we have already proposed in the paper [] that allows to ensure the operation of Intelligent transportation system even outside the technological infrastructure figure 1, the present paper develops the part related to the updating of the graph by the new values of the arcs (yellow part of the figure 1) , which represent the real state of the roads, this work is based on NLP techniques belonging to the field of machine learning and big data architecture.

5. 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, twitter and Facebook, through their APIs.

5.2 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.

5.2 Facebook Grraph 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 [25][26].

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

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. PROPOSED WORK

We present our suggested multi-level pattern for a novel approach to traffic incident detection based on social big data, apache spark, and machine learning as mentioned in figure 3. this architecture has six basic layers: obtaining information and storage layer, data pre-processing layer, feature extractor layer, tweet/comment filtering layer, event detection layer, and results visualization.

Beforehand, when data are collected on json format, they get stored as objects[27] into a mongodb database[28], we process 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[29], bug of words (bow) serves also for the same objective[30],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.

We used three models for event detection, each of which used three distinct supervised classification algorithms. we then used the four frequently used evaluation metrics, precision, accuracy, recall, and f-score, to evaluate the models and then choose the best algorithm 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 data frame, it is based on the rdd (resilient distributed datasets)[31] which allow for higher performance and higher speed of computing, the main steps of our system are described below:

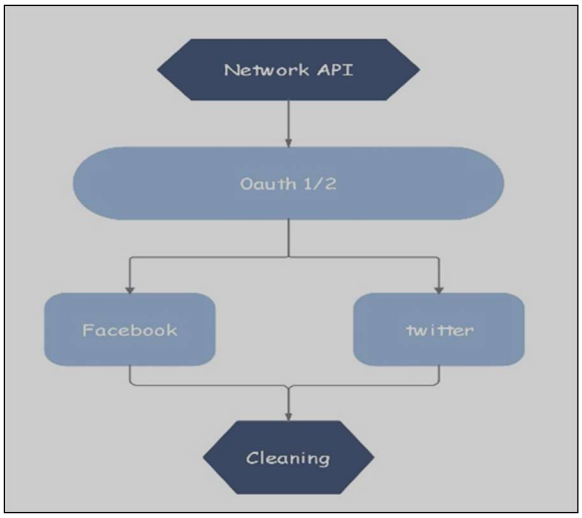

Figure 2: Social Data Collection Model

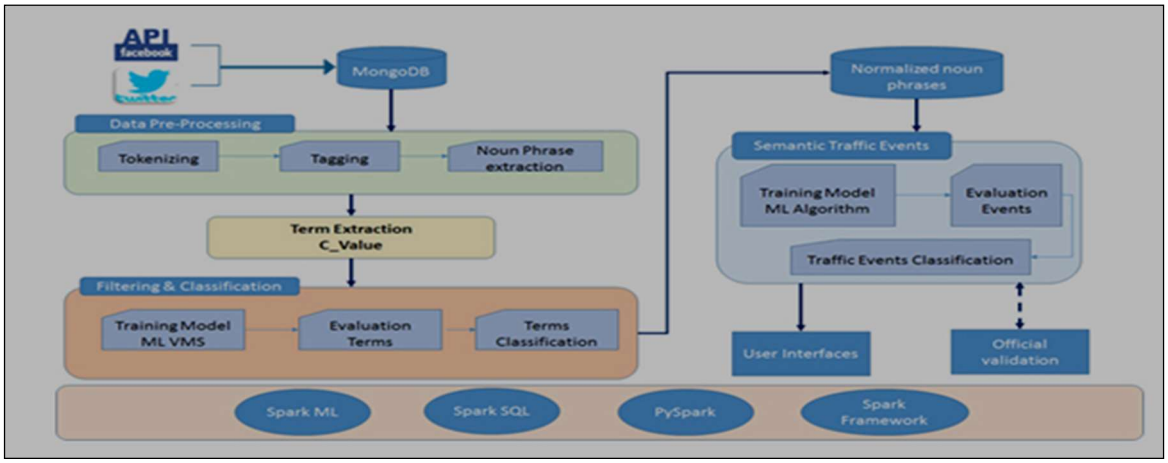

Figure 3: Traffic Events Detection Approach based on "Social Big Data "and Machine Learnings Technique

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6.1 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 geolocalization 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 august 1, 2020,We opted for NoSQL databases over relational databases because our data required scalable and flexible Schema based storage. The comments/tweets are stored in MongoDB, a document-oriented database designed for storing and managing huge collections of documents. The JSON objects collected from our two sources are used to update the MongoDB database. The 'made at,' which reveals the time the tweet was posted, and the 'full text,' which offers the message content, are two attributes of the gathered object. After that, we looked for and deleted duplicate comments/tweets.

6.2 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[32], 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 ('أ') ,'إ' ,'آ' and normalized to bare Alif ('ا'), this is done with all Arab letters written indifferent forms. We have to drop the stop words included on gathered texts by using the Stop-Words list provided through Toolkit (NLTK) [33]. 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, in this work we utilize, additionally to Apache Spark OpenNLP, AraVec[34] 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.

.

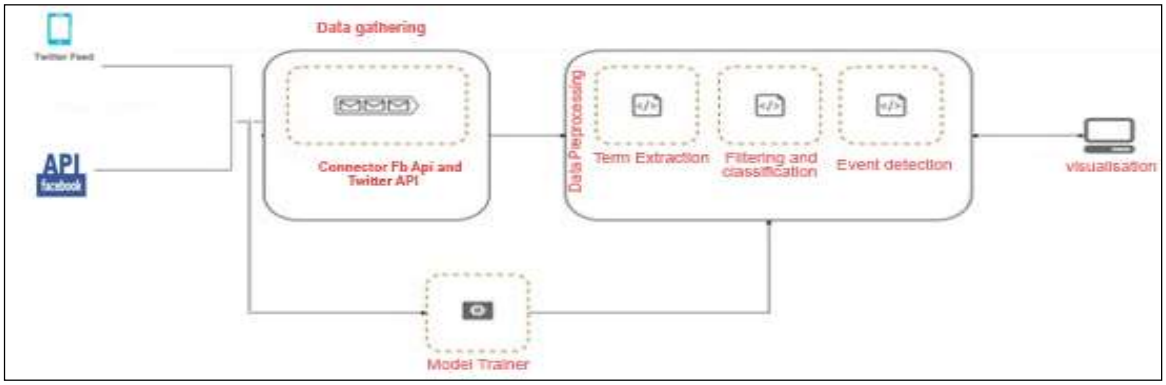

Figure 4: Simple Pipeline Design for NLP Procedure to Detect Related Transportation Events

6.3 Term Extraction and Selection

We employ the Term Frequency/Inverse Document Frequency TF/IDF to construct this specific stage, which is based on the spark machine learning techniques. The primary goal of this technique is to determine the significance of a word within a document (comments/tweets). The I DF (tr, doc) is calculated by calculating the number of times the term or its synonyms appear in the document. The function T f (tr, doc) is the recurrence of the appearance of term tr in a document doc. The following equation is used to compute the IDF value:

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

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

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

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 "Feature" which constitutes the input for the filtering and classification stage 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.

6.4 Filtering and Classification of Terms

The collected tweets are related to various topics, it's needed for filtering only data related to traffic topic, for this a specific filter algorithm that contains spark machine learning package is used. We divided the manually labeled data into training (80%) and testing (the remaining 20%). After that, we use the Nïve Bayes, SVM, and logistic regression (LR) methods to develop and train the model. On the training set, the models are trained. On the testing set, we evaluate the algorithms and choose the best one. The trained classifier is used to test precision, accuracy, recall, and F-score using popular statistical metrics. To clarify the meaning of these measurements, we refer to traffic-related tweets as positive class and none-related tweets as negative class. The following four classes are used in these metrics: True Positive (TP) refers to positive tweets that were successfully anticipated as positive, True Negative (TN) refers to negative tweets that were correctly forecasted as negative, and False Positive (FP) refers to tweets that were incorrectly forecasted as 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.

(Eq3) $$\mathrm{acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}}$$

(Eq4) $$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

(Eq5) $$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

(Eq6) $$\mathrm{F}(\beta) = (1+\beta 2).\left(\frac{\mathrm{PPV.TPR}}{\beta 2\cdot \mathrm{PPV+TPR}}\right)$$

The documents' vector representation allows us to view sub-areas, each of which represents a specific topic connected to our primary topic "traffic topic." By comparing our term class (subject) to the predefined themes inside the vector space, we can train our model to detect sentences or words related to our topic. However, we must assess the training model's output using a set of measures like as accuracy, precision, rating score, and recall. These measurements help us to verify that the classifier's results are accurate and relevant to the traffic topic. As a result, two classifications will be identified: positive terms and negative TERMS PT) and negative terms (NT).

6.5 Event Detection

In this level we set up a classifier then we train it using naïve bayes, logistic regression and svm. the events classifier is trained after labeling manually part of the filtered data from the last stage into eight event classes, which are weather, fire, social events, traffic condition, roadwork, road damage, accident, and road closures, these classes have positive and negative comments/tweets about the traffic condition. for some of theme we have to consider all kind of data (positive or negative), whereas for the

rest we consider only data which can affect negatively the road conditions. we've noticed that some event categories receive a lot of more tweets than the others. basing on the number of tweets, we categorize them into small-scale and large-scale events. traffic conditions, roadwork, road damage, accidents, 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

7. RESULTS AND DISCUSSION

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.

In this section, we describe the results of a case study on filtering and classifying various transportation-related events in Casablanca using social big data (Facebook/Twitter) analysis. For data filtering, we examine the performance of three commonly used algorithms: SVM, NB, and Logistic Regression

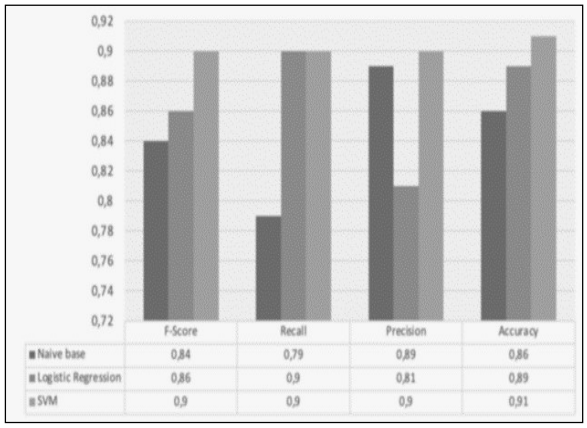

Figure 5: Result Analysis for Comments/Tweets Filtering

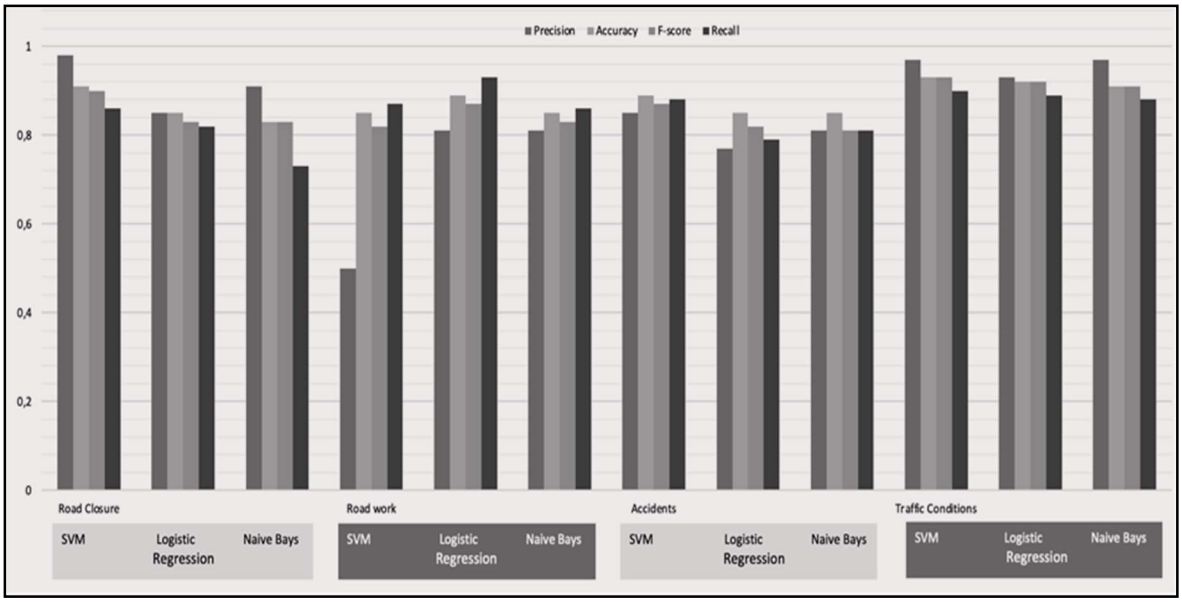

Figure 6: Evaluation results for events classification Accuracy, Precision, Recall and F-score

Journal of Theoretical and Applied Information Technology
15th June 2022. Vol.100. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

According to the figure5, 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%.

As mentioned in figure 6, 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 the useful among them according to our goals.

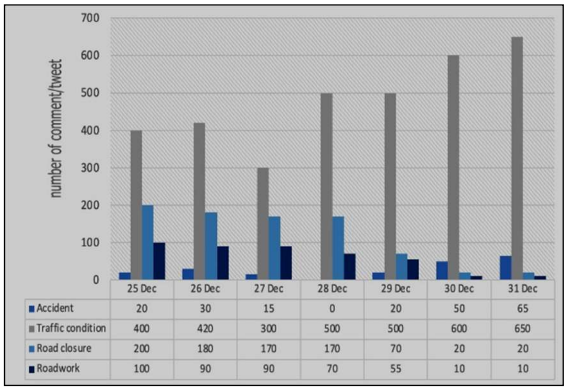

Figure 7: Number of Detected Large Scale Events.

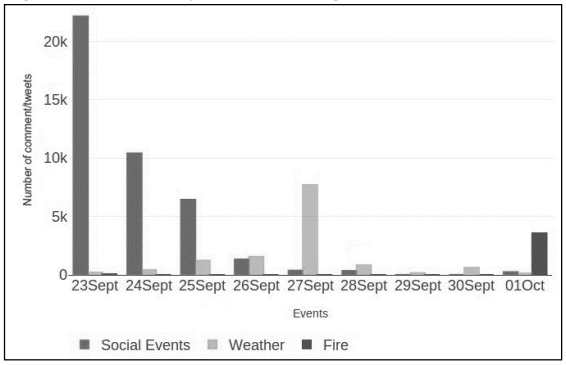

Figure 8: Number of Detected Small Scale Events.

We divide data in the gathered dataset into two events categories (figures. 6 and 7) based on the number of events and the type of event, thus in our instance, we have two categories: small scales such as road closures, road work, traffic conditions, and accidents. We have road damage, accidents, and traffic congestion on a massive scale. To determine the reliability of each event, we count the number of comments/tweets containing these terms within the context of the event location and the date of publication, resulting in the number of events each day. To detect accident events, we extract and capture events or vocabulary which include " كسيدة في الطريق بين شارع 55 و 33 " 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. Figure 7 shows that the number of comment/ tweets telling about accident increases at the working time between 7h30 to 9h:30, 12h to 12h30, 14h to 14h30 and 18h to 19h

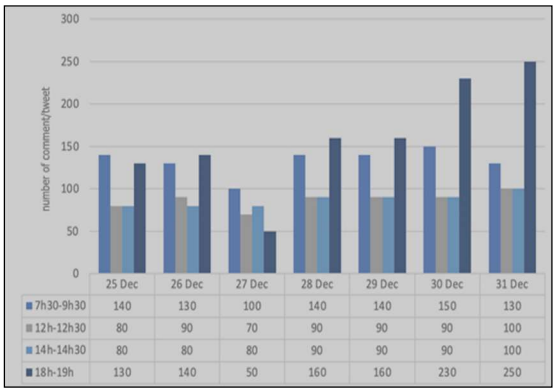

Figure 9: Number Comments/Tweets During Different Time Interval per Day and the Type of Event.

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. The presented results confirm the effectiveness of our approach for filtering and detecting traffic events based on tweets as well as the time and the location of each event.

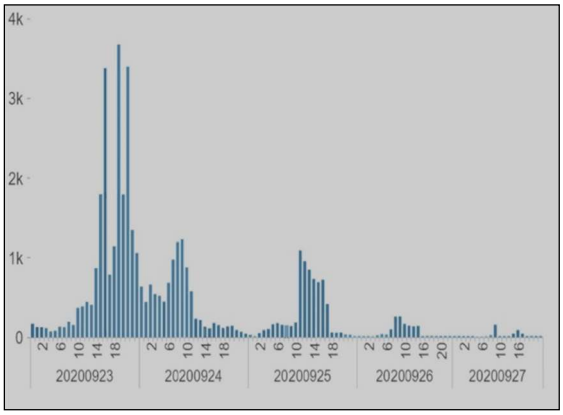

Figure 10: The number of comments/tweets per hour for the top 'Social Event'

8. CONCLUSION

Despite the increasing velocity of data offered by social networks, detecting road traffic events remains challenging for transportation applications. However, the traffic event detection has been automated in the previous case study using machine learning techniques based on the NLP domain and the Apache Spark big data platform. The way Moroccan language is written was such a serious challenge due to the fact that it includes many languages and many different 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.

Furthermore, a classifier was trained to filter data to determine whether it was relevant or irrelevant to a transportation topic. We used the following three machine learning algorithms: Nave Bayes, SVM, and LR, and then we trained the other classifiers to detect the occurrence of various transportation-related events in Casablanca region. Following that, all transportation-related events are classed; we acquire information on the event's category, location, and timing for each event, and we confirm this information by reviewing official sources such as official websites or electronic newspapers. The proposed method's capacity TO detect real transportation-related events in real time has been demonstrated experimentally.

This approach of near-real time detection of the road stat used to ameliorate and complement the intelligent transportation framework proposed in the study [7],for instance, we can rely on the near-real state of roads information so as to update the graph's values in the second study, where the graph is used as a model to solve the salesman problem and their values represent the cost of arcs (paths, roads),that connects different edges (cities, destinations points),these effective cost values allow, henceforth, a high degree of optimization in term of finding the best itinerary to be taken.

9. PERSPECTIVES

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:

[1] . Carel, L., *Big data analysis in the field of transportation*. 2019, Université Paris Saclay.
[2] . Tseng, Y.-y., W.L. Yue, and M.A. Taylor. *The role of transportation in logistics chain*. 2005. Eastern Asia Society for Transportation Studies.
[3] . Barreto, L., A. Amaral, and T.J.P.M. Pereira, *Industry 4.0 implications in logistics: an overview.* 2017. 13: p. 1245-1252.
[4] . Nesmachnow, S., S. Baña, and R.J.E.E.T.o.S.C. Massobrio, *A distributed platform for big data analysis in smart cities: combining intelligent transportation systems and socioeconomic data for Montevideo, Uruguay.* 2017. 2(5).
[5] . Galati, S.R., *Funding a Smart City: from concept to actuality*, in *Smart Cities*. 2018, Springer. p. 17-39.
[6] . AlEnezi, A., Z. AlMeraj, and P. Manuel. *Challenges of IoT based smart-government development*. in *2018 21st Saudi Computer Society National Computer Conference (NCC)*. 2018. IEEE.
[7] . MOUAMMINE, Z., et al., *innovative architecture based on big data and genetic algorithm for transport logistics optimization.* 2020. 98(17).
[8] . Mouammine, Z., et al., *Big Data with Distributed Architecture Using Genetic Algorithm in Intelligent Transport Systems.*
[9] . Lavanya, B.M.a.K., *Social Media Data Analysis for Intelligent Transportation Systems*, in *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. 2020. p. 1-8.
[10] .Conway, M., M. Hu, and W.W.J.Y.o.m.i. Chapman, *Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data.* 2019. 28(1): p. 208.
[11] .Rahman, A. and M.S. Hossen. *Sentiment analysis on movie review data using machine learning approach*. in *2019 International Conference on*

Bangla Speech and Language Processing (ICBSLP). 2019. IEEE.

[12] .Polignano, M., et al. *Alberto: Italian* BERT language understanding model for NLP challenging tasks based on tweets. in 6th Italian Conference on Computational Linguistic*s, CLiC-it 2019*. 2019. CEUR.

[13] .Tachicart, R., K.J.J.o.K.S.U.-C. Bouzoubaa, and I. Sciences, *Moroccan Arabic vocabulary generation using a rule-based approach.* 2021.

[14] .Preda, S., S.-V. Oprea, and A.J.S. Bâra, *PV forecasting using support vector machine learning in a big data analytics context.* 2018. 10(12): p. 748.

[15] .Granik, M. and V. Mesyura. *Fake news detection using naive Bayes classifier*. in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. 2017. IEEE.

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

[17] .Shafiq, M., et al., *Data mining and machine learning methods for sustainable smart cities traffic classification: a survey.* 2020. 60: p. 102177.

[18] .Lan, S., et al., *Trends in sustainable logistics in major cities in China.* 2020. 712: p. 136381.

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

[20] .Rettore, P.H., et al., *Road data enrichment framework based on heterogeneous data fusion for its.* 2020. 21(4): p. 1751-1766.

[21] .Banerjee, S., et al., *Units of information on dynamic message signs: a speed pattern analysis.* 2019. 11(1): p. 1-9.

[22] .Ma, Z., et al., *Analyzing drivers' perceived service quality of variable message signs (VMS).* 2020. 15(10): p. e0239394.

[23] .Sandt, A., et al., *Using Agency Surveys and Benefit–Cost Analysis to Evaluate Highway Advisory Radio as Regional Traveler Information and Communication Tool.* 2017. 2616(1): p. 81-90.

[24] .Mfenjou, M.L., et al., *Methodology and trends for an intelligent transport system in developing countries.* 2018. 19: p. 96-111.

[25] .TheArabWeakly. *Morocco, China agree on financing for $10 billion tech city*. 2017 [cited london 03.07.2021]; 99:[Available from: https://thearabweekly.com/morocco-china-agree-financing-10-billion-tech-city.

[26] .Statistica. *Distribution of tweets per user per day from the Middle East and North Africa in March 2016, by country*. 2016 Aug 26, 2020; 2017:[Available from:

[27] .Marr, B. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. 2021 [cited 2021 09-08-2021];

[28] .Kumar, J. and V. Garg. *Security analysis of unstructured data in NOSQL MongoDB database*. in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*. 2017. IEEE.

[29] .MongoDb. *MongoDb website*. 2021 [cited 2021 17-07-2021];

[30] .Kim, S.-W., J.-M.J.H.-c.C. Gil, and I. Sciences, *Research paper classification systems based on TF-IDF and LDA schemes.* 2019. 9(1): p. 1-21.

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

[32] .Jonnalagadda, V.S., et al., *A review study of apache spark in big data processing.* 2016. 4(3): p. 93-98.

[33] .Ali, M.A.J.I.J.o.A. and A. Sciences, *Artificial intelligence and natural language processing: the Arabic corpora in online translation software.* 2016. 3(9): p. 59-66.

[34] .Loper, E. and S.J.a.p.c. Bird, *Nltk: The natural language toolkit.* 2002.

[35] .Mouammine, Zakaria, Bourekkadi & AL, *parallel optimization of routing and improvement of the energy consumption of an electric vehicle, education excellence and innovation management: volume issue page 4380-4386* 2020.

[36] ..S. Ennima et al., Innovation of the photovoltaic system, general and maximum power point tracking device for smart cities, journal of theoretical and Applied Information Technology, 15th June 2021, Vol. 99. No. 11

[37] .El Imrani, O., Ben Messaoud Layti, M., Bourekkadi, S., Boulaksili, A., Kabbassi, I. , *Optimization of the International Trade Activities in the Period of COVID-19 by Proposing an Algorithm , Studies in Systems, Decision and Control*, 2021, 358, pp. 187–1

[38] .Manar Kassou et al. , *Digital transformation in flow planning: the case of container terminals at a smart port , Journal of Theoretical and Applied Information Technology* , 15th May 2021. Vol.99