Chapter 2
Sentiment Analysis of Arabic Tweets for Road Traffic Congestion and Event Detection

Ebtesam Alomari, Rashid Mehmood, and Iyad Katib

2.1 Introduction

More recently, Twitter has become a popular social platform to share traffic information. Mainly, Twitter can provide information about future events, the causes behind certain behavior, anomalies, and accidents, as well as the public feelings on a matter. Furthermore, there are specific, and official Twitter accounts created to report on traffic conditions and events in particular cities. These accounts generate useful sources of information for the followers. Hence, there is an enormous amount of traffic updates and information available in different Twitter accounts and can be freely obtained via the easy-to-access APIs [1].

Several researches have been proposed to monitor road traffic in different countries by analyzing text from different languages such as English and Chinese. However, the difficulty of performing such analysis in Arabic social media lies in the fact that the dialectical Arabic is used more than the formal Modern Standard Arabic (MSA), which produce new challenges for Arabic text classifications and Sentiment Analysis (SA) [2]. To the best of our knowledge, none of the existing works about sentiment analysis on Saudi dialect tweets have focused on traffic condition. Moreover, the existing analysis approaches for Arabic event detection did not focus on road traffic in Saudi Arabia. Further, they did not apply big data

E. Alomari (✉) · I. Katib
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: EAlomari0011@stu.kau.edu.sa; iakatib@kau.edu.sa

R. Mehmood
High Performance Computing Center, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: RMehmood@kau.edu.sa

technologies to properly handle such huge amounts of social data which required high processing speed, large storage, and other challenges.

Currently, road traffic congestion is one of the biggest problems in Saudi Arabia especially in large cities like Jeddah. Jeddah city is the second largest city in Saudi Arabia and arguably the most congested one. Further, Makkah is the Islam's holiest city, where millions of Muslims visit. The increasing number of vehicles and an enormous number of pilgrim visitors all year round have increased accidents and traffic jams in many major roads in this city. Moreover, the Kingdom accounts for over 40% of all active Twitter users in the Arab region [3]. By 2016, the number of Twitter users in Saudi Arabia had reached 4.99 million [4]. Hence, Twitter presents an excellent opportunity for extracting useful information. This raises the need for proposing a novel approach to analyze Arabic dialectical social data to monitor road traffic in Saudi Arabia.

In this paper, we extend our previous work [5] about analyzing and extracting traffic congestion information from Arabic tweets. In our previous work, we analyzed only negative tweets that refer to traffic jam and congestion where we designed the search queries to fetch tweets that contain specific negative traffic-related keywords. Subsequently, we extracted the traffic causes and the top congested roads and streets in Jeddah city.

In this work, we collect all traffic-related tweets regardless of the type (negative or positive). We fetch tweets about traffic in Jeddah and Makkah during Ramadan, which is the Islam's month of fasting. We chose this period to study the impact of this month on road traffic because in this month the traffic behavior and the road traffic rush hours change significantly. The main objectives and contributions of this paper can be summarized as follows:

– Improve our previous methodology by applying normalization on the extracted Arabic tokens.
– Provide a mechanism to detect events that could affect the traffic condition.
– Propose an approach for sentiment analysis to classify a driver's feeling and emotions.

Sentiment classification is one of the areas in which "big data" requires processing. Thus, we have built our approach on SAP HANA, which is an in-memory processing platform that can help to improve both the performance and the quality of the results. We analyzed the data by applying a lexicon-based approach. We have built lexicons (dictionaries) for Arabic and Saudi dialect words. The dictionaries include the most common words regarding traffic condition. The main goal is to classify traffic-related tweets into one of four sentiment classes (Strong positive, Positive, Strong negative, and Negative).

The rest of the paper is organized as follows. Section 2.2 reviews the related works. Section 2.3 illustrates the methodology. Section 2.4 discusses the results. Finally, we draw our conclusions in Sect. 2.5.

2.2 Literature Review

2.2.1 Transportation and Smart Cities

Traditional approaches for traffic measurement have relied on sensors that are buried under the road (such as inductive loops) or installed on roadside [6]. Additionally, many traffic monitoring systems have been proposed to detect road congestion using video [7] or image [8] processing technologies. However, these approaches require sensors and other equipment such as cameras and thus the deployment and maintenance are costly.

Several approaches have been proposed, particularly during the last decade, to use vehicular ad hoc networks (VANETs) for monitoring traffic [6, 9, 10], in general, and for specific purposes, such as for traffic coordination and disaster management [11–13]. Simulations have also been playing a key role in transportation planning and control [14]. A number of works on operations research related to transportation in smart cities have also been proposed, e.g., car-free cities [15], intelligent mobility [16], big data in transport operations [17, 18], prototyping in urban logistics [19], and autonomic transportation systems [20–22]. Furthermore, Alomar et al. [23] visualized traffic incidents in the city of Riyadh for the 2013–2015 timeframe. However, they did not work on social data. They get the data from the General Directorate of Traffic (GDT). Other researchers study road traffic crashes in Pakistan during Ramadan [24]. They also analyzed structured data from formal sources.

2.2.2 Event Detection from Social Media

Several approaches have been proposed to detect events from social data in different languages.

Kurniawan et al. [25] conducted experiments to classify real-time road traffic tweets using data mining. They collected real-time data about Yogyakarta Province, Indonesia using Twitter Streaming API. Additionally, they compared classification performance of three machine learning algorithms, namely Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT). However, they only classified tweets into the traffic or non-traffic categories. Similar work is proposed by D'Andrea et al. [26]. They suggested an intelligent system, based on text mining and machine learning algorithms. They collected real-time tweets of several regions of the Italian road networks and then assigned the appropriate class label to each tweet, as to whether the tweet is related to a traffic event or not.

Ribeiro et al. [27] analyzed tweets to detect traffic events in Belo Horizonte, Brazil. They created a set of place names, called GEODICT. Subsequently, they detected the locations and streets names by using string matching technique by searching for substrings from the tweet that can be detected in GEODICT. Wongcharoen and Senivongse [28] built a congestion severity prediction model to

predict traffic congestion severity level. However, like previous approaches [25, 27], the tweets are fetched only from particular accounts.

Hanifah et al. [29] filtered tweets using SVM to detect traffic congestion in Bandung, Indonesia. Also, they extracted the information of location, time, date, and image. For information extraction, they applied a rule-based approach, which is based on the structure of the sentence. However, they did not detect traffic-related events. Gu et al. addressed this limitation [30]. They have collected historical and real-time tweets about traffic in Pittsburgh and Philadelphia, Metropolitan. They used a dictionary of relevant keywords and their combinations that can indicate traffic condition.

Moreover, D'Andrea et al. [26] collected real-time Italian tweets and classified them into traffic and non-traffic tweets. Alifi and Supangkat [31] suggested approaches for extracting location information. Additionally, they extracted valuable information from real time including traffic condition, congestion causes, weather condition, and time of occurrence. However, researchers in [26, 29, 30], and [31] did not perform sentiment analysis. Additionally, none of them applied big data technologies in their proposed methods. Suma et al. [32, 33] have analyzed Twitter data to detect events related to road traffic and other topics for smart cities planning purposes. Their focus is on the use of big data platforms to analyze large amounts of tweets about the London city. However, they did not perform sentiment analysis. Moreover, in our previous work [5] we used SAP HANA to detect road traffic conditions in Jeddah city. However, we did not perform SA.

Several approaches have been proposed to detect events from Arabic social data. AL-Smadi and Qawasmeh [34] used an unsupervised rule-based technique to extract events about technology, politics, etc. In [35], the researchers detect events related to disasters, sports, arts, crime, politics, and elections. Other researchers classified real-time tweets to detect high-risk floods [36]. Moreover, researchers in [37] annotated Arabic events related to politics and election. Furthermore, Alsaedi and Pete [38] proposed a framework for detecting disruptive events from Arabic tweets. They extended their work and suggested an integrated event detection framework related to the riots events [39]. However, none of these studies focused on traffic events.

2.2.3 Arabic Sentiment Analysis

The existing work about Arabic sentiment analysis (not specific to transportation) can be classified into lexicon (dictionary) based, ML-based, or hybrid. Researchers in [40–42] applied a hybrid approach for Jordanian dialect. On the other side, there are some studies based on machine learning for Modern Standard Arabic (MSA) [43], Egyptian dialect [44], and Jordan dialect [45]. Furthermore, researchers in [46, 47] proposed lexicon-based Arabic SA, but they are not proposed for Saudi dialect.

Few studies have applied SA to Saudi dialect. Aldayel and Azmi proposed hybrid (SVM and lexical) classifier [2]. However, they only performed two-way (positive, negative) classification. Moreover, the Saudi dialect lexicon has been developed

in [48]. But, it is domain specific (restaurants reviews). Al-twairesh proposed AraSenTi-tweet [49] corpus for sentiment analysis. It is available online for the research community. Even though the corpus annotated manually, they extracted from a large dataset that contains Arabic tweets. Most of the existing words in their lexicon are not useful in our case (traffic detection). Further, some of them do not belong to the Saudi dialect.

From the above discussion for the literature review, we found that the existing Arabic sentiment lexicons are either not supporting Saudi dialect or not efficient to be used in traffic detection domain. Therefore, there is a need to create a new sentiment lexicon to classify the traffic-related tweets.

On the other side, big data processing technologies provide great opportunities for addressing transportation problems for which traditional approaches are not competent. To the best of our knowledge, none of the existing work about event detection from Arabic social data has used big data platforms and technologies to address the complex processing and analytics tasks on such big data. Therefore, our text classification technique will be built on SAP HANA, which is an in-memory processing platform offering groundbreaking performance.

2.3 Methodology

Figure 2.1 illustrates the workflow of tweets acquisition, processing, and analytics. We have built our approach on SAP HANA, which is developed by SAP SE. It is the integration of transactional and analytical workload within the same database
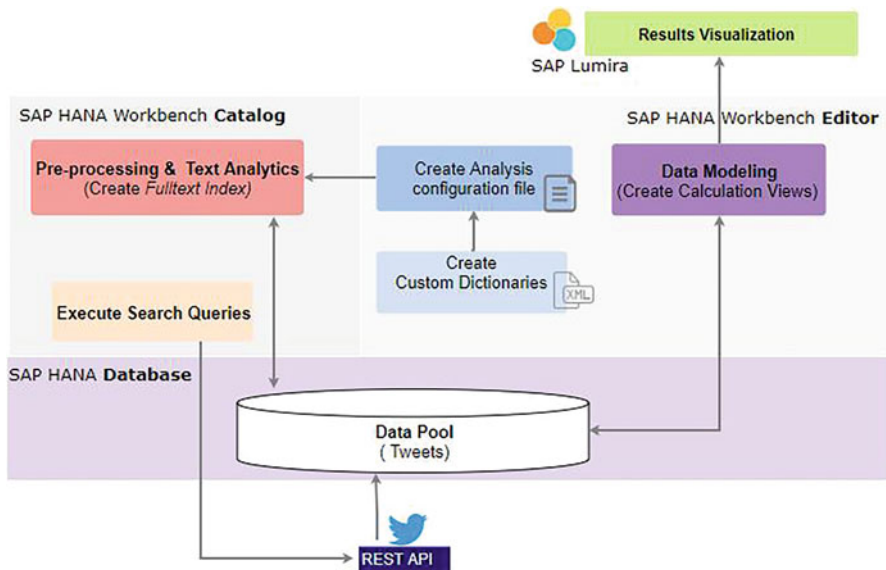

Fig. 2.1 Overview of the main implementation steps

management system [50]. Further, SAP HANA Extended Application Services (SAP HANA XS) provides the SAP HANA *Web-based Development Workbench* that supports developing entire applications in a Web browser without the need to install any development tools. SAP HANA Web-based Development Workbench includes i) Catalog and ii) Editor tools [51].

Catalog enables developing and maintaining SQL catalog objects in the SAP HANA database. It also supports creating tables, executing SQL queries, and creating a remote source to collect data. Additionally, *catalog* supports text analysis and text mining. Moreover, *Editor* enables data modeling, which is an activity of creating information view. This information views can be used for reporting and decision-making purpose. SAP HANA supports a great information view, which is a calculation view. The data foundation of the calculation view can include tables, column views, analytic views, and calculation views. Also, it enables creating joins, unions, aggregation, and projections on data sources.

2.3.1   Tweets Collection

We collected tweets about traffic in Jeddah and Makkah during Ramadan (17 May– 14 June), 2018. We generated a list of Arabic keywords related to road traffic and transportation. We also searched for the most popular Twitter accounts that tweet about traffic conditions in Jeddah and Makkah cities. We have used the collected list of twitter accounts and Arabic keywords to write a large number of queries.

Search queries were executed in SAP HANA Workbench *Catalog* to collect historical tweets using twitter REST search API. Unlike streaming API that enables fetching real-time tweets, the REST API allows us to query historical tweets with locations and keywords simultaneously. REST API supports geocode parameter to restrict query by a given location using "latitude, longitude, radius." Thus, when executing the queries, the search API will first attempt to search for tweets which have lat/long within the queried geocode. But not all tweets are geotagged because some users disable location service in their smartphones. In this case, Tweet's location information will be detected from the location data in the user's profile.

However, if the user did not add information about the city and county in his/her profile, "Country" and "Place_name" fields would be empty. To handle this issue and fetch the non-geotagged tweets, we re-execute all queries after adding the city name and without specifying a location to collect all traffic tweets that include the city name. However, there are still some tweets that are not included in our analysis because they are not geotagged and not carrying location information. We created a table to store the retrieved tweets in SAP HANA databases. The created table includes several attributes such as "UserId," "Tweet," "UserName," "CreatedAt," "Latitude," "Longitude," "Country," and "Place_name."

2.3.2 Pre-processing and Analysis Configuration

SAP HANA supports text analysis for different languages including Arabic. They used the pre-processor server to extract and classify unstructured text into entities and domains by applying linguistic and statistical techniques [52]. To analyze the text in SAP HANA, there is a need to create full-text indexing on the text column with specifying the type of analysis configuration and setting TEXT ANALYSIS parameter "ON" and this results in a new table "$TA__<index name>". This table will include linguistic or semantic analysis results.

SAP HANA supports three main types of text analysis configurations, which are [53]:

– Linguistic Analysis: supports natural language processing.
– Entity and Fact Extraction: enables named entity extraction, sentiment analysis, public sector events, and enterprise facts. It named EXTRAC-TION_CORE_VOICEOFCUSTOMER.
– Grammatical Role Analysis: enables functional syntactic roles in the sentence, such as subject or object. It supports English language only.

In this work, the data are analyzed based on "Voice Of Customer" (VOC) analysis configuration. We have selected this type of text analysis configuration because it supports handling entity extraction, fact extraction, and sentiment analysis. Further, it enables tokenization, which means it decomposes the phrase or sentence into tokens. Unlike "Linguistic analysis" configuration that extracts every word in the text, VOC extracts only basic entities from the text and entities of interest including a person, address, organization, URLs, and other common terms. The token type is stored in TA_TYPE field.

To use the default configuration, developers simply need to include VOICEOF-CUSTOMER parameter in a query. However, the standard configuration doesn't suffice to the requirement especially with the Arabic language. Further, the default normalizer is not efficient. Thus, we need to customize keywords in new dictionaries and include them in a modified configuration file.

Custom Dictionaries

We noticed that the standard text analysis in SAP HANA using the VOICEOF CUSTOMER-configuration does not suffice where not all Arabic tokens are classified under the right token type. Therefore, we need to add a custom dictionary for unknown terms in the SAP HANA system and then create a new configuration file. We created our own dictionaries because none of the existing dictionaries for Saudi dialect are designed to be used for road traffic condition detection. The created lists of custom dictionaries were used to create a new configuration file for analysis using SAP HANA Web-based Development Workbench. Then, the generated configuration file was used to create the fulltext index on "Tweets"

column to split the text into tokens and specify the token type based on the created dictionaries.

We created several custom dictionaries, which help to improve tokenization, normalization, and entity type extraction. The main dictionaries are as follows:

– Transportation: includes the collected Arabic keywords about transportation (such as سيارة, ساهر, إشارة).
– Makkah Streets/Jeddah Streets: contain the names of streets and roads names.
– Places: includes the keywords referred to places names like Mosque, Restaurant, and Mall.
– Religion: contains the synonyms of words related to fasting and the activities during Ramadan month (e.g., افطار, عمره, التراويح).
– Sentiment: includes a list of Arabic and Saudi dialect sentiment words and expression.
– Events types: contains the common words representing events types and list of their corresponding synonyms.

Tokenization, Normalization, and Entity Extraction.

To analyze the tweets in SAP HANA, we need to create a full-text index on "Tweet" column. Creating the index requires executing SQL statement, which will lead to creating a new table containing the tokens and named entity extraction results. The created table will include the following:

– TA_Token: contains the list of tokens extracted from the tweets.
– TA_Type: refers to the entity type.
– TA_Normalized: stores a normalized representation of the token.

The created custom dictionaries enable identifying a standard name for each entity. The TA_Type field can contain built-in type (e.g., NOUN_GROUP) or one of the types that are specified in our newly created dictionaries, i.e., Jeddah_Street. Moreover, the normalization process is very important especially for Arabic text where some letter has different representation. For instance, "Alif" has four forms (أ, ا, إ, آ), "Yaa" has two forms (ى, ي), and "Haa" has two forms (ه,ة). SAP HANA supports case normalization by converting the initial letter of a word to upper or lower case. However, this type of normalization is not relevant to languages that do not distinguish between upper and lower case such as Arabic. So, we modified the analysis configuration to represent the normalized form of the entity as specified in our custom analysis dictionaries. For example, "زحمة" and "المدينة" will be normalized to "زحمه" and "المدينه" where "TAA MARBUTAH/ة" was replaced with "HAA/ه."

2.3.3 Tweets Analysis

Location Extraction

Generally, there are two types of location information: (i) Latitude/longitude coordinates of the locations where users posted the tweets and (ii) Location name referred in tweet texts. We specified either coordination information or cities name in our search queries to force them to retrieve only tweets posted in our targeted cities. Further, to extract specific location information such as streets name from the text, we used the Entity Extraction feature in SAP HANA. However, the existing entity extractor with default configuration did not detect all the places names. So, we created our own dictionaries for the main streets/roads names and then we included them in the modified configuration file. We used OpenStreetMap[1] to create a list of streets and roads names in Jeddah and Makkah. When we run the analysis query (create full-text index), the places name will be extracted from the text and stored in the analysis results table.

Traffic Events Detection

We created a dictionary containing a list of words representing the road traffic events. We took into account the following events:

– Accident (حادث).
– Fire (حريق).
– Roadworks "أعمال الطرق" including maintenance (صيانة) and construction (بناء).
– Weather condition "الطقس" such as rain (مطر) and storm (عاصفة).
– Other events that could affect the traffic including sports (رياضة) events and social events (e.g., festival "مهرجان").

We expand the dictionary by adding a list of corresponding synonyms under each event type. Consequently, each type of traffic event is extracted taking into account the set of relevant words. For instance, accident "حادث" associated with words like "صدم," and maintenance "صيانه" associated with words like "ترميم" or "اصلاح." To clarify, during the tokenization and entity extraction phases, each token will get a Token_Type based on our custom dictionaries where our event detection technique relies on matching synonyms with terms available on the tweet. For instance, the following tweet contains the word "حريق," and thus the extracted event type will be "fire." We consider the fires as traffic-related events even though it is not a vehicle fire because it may effect on the traffic condition and cause congestion.

Example: "@JeddahNow: حريق ضخم في متاجر #اكسترا على شارع التحلية ، مع تواجد كثيف لفرق الدفاع المدني ، سنوافيك بالمستجدات لاحقاً مباشر #جدة".

---

[1] http://openstreetmap.org

Translation: "@JeddahNow: Live #Jeddah | A huge fire at # Extra stores on Tahlia Street, with an intensive presence of the Civil Defense teams, we will update you about the status soon."

Sentiment Analysis

The literature review suggests two approaches for building a lexicon: manual construction by experts or automatic construction. Although automatic lexicon construction from a seed of words is faster and required less human effort, there are weaknesses regarding accuracy and robustness due to the lack of human supervision. Thus, in this work, we followed a lexicon-based approach that relies on a manually constructed dictionary. We built lexicons for Saudi dialect words that related to traffic condition. We created a list of strong positive words (e.g., اسرع "Faster"), positive words (e.g., سالك "no traffic jam"), negative words (e.g., بطيء "Slow"), and strong negative words (e.g., وفاه "Death"). Then, we expanded the lists by adding synonyms.

After that, we included the created custom dictionaries in the analysis configuration file. When we created a full-text index, the analyzer simply splits each word in the tweet, normalize it using our dictionaries, then classify each token in the tweet into one of the four categorized. Subsequently, we created a calculation view to classify the tweets. Each tweet will be scored based on the number of the tokens from each sentiment class and on how many times these words occurred in the text. Subsequently, the tweets are classified appropriately based on the calculated score.

2.4   Results and Dissection

SAP offers a data visualization tool for reporting on top of SAP HANA, named SAP Lumira.[2] Figure 2.2 shows the percentage of tweets at different time of day. The chart in Fig. 2.2a shows that most tweets about traffic in Jeddah are posted during the night. The highest tweeting time is at 22. The percentage of tweets is started decreasing after 3 and the lowest tweeting time is at 8. The results are reasonable where the business hours during Ramadan are changed, and people used to go to the markets and restaurants before Iftar in addition to that they usually go shopping after Al-Taraweeh prayer Additionally, during Ramadan, the work hours are changed, and most employees in public sector and private companies work from 10 am to 3 pm.

On the other side, Fig. 2.2b shows that the percentage of tweets about traffic in Makkah is always high except for the period between Al-Fajr prayer and Al_Dhuhr
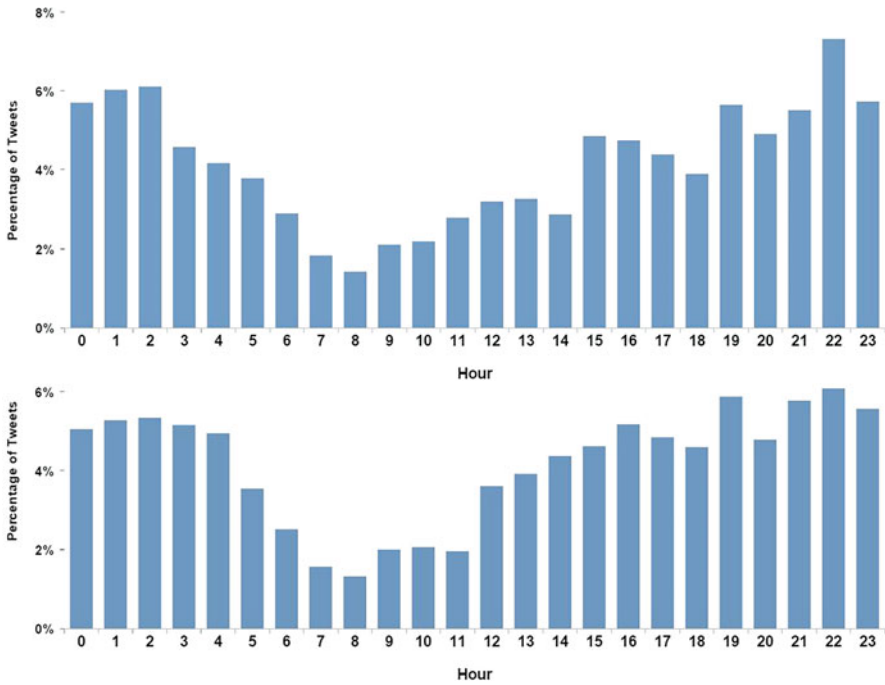
Fig. 2.2 Percentage of Tweets on different time. (**a**) Jeddah. (**b**) Makkah

prayers (5–12). Like the tweets about Jeddah, the number starts decreasing after Al-Fajr prayer where most people used to sleep at this time.

Moreover, we filtered the collected tweets to show only the tweets messages that contain street/road names. After that, we drew a chart to represent the top mentioned street/road. However, we noticed that the number of tweets messages that contains place name is not very large. The main reasons that could explain that are (i) the limit in the characters number in Twitter, (ii) people may post a message to reply to another tweets or participate in a hashtag about specific events, which don't required re-mentioning the name of the place, (iii) the tweets that describe feelings or emotions usually do not contain a specific place name.

As shown in Fig.2.3a, the most mentioned names in the collected tweets about traffic in Jeddah are Prince Sultan St., Altahliah St., King Abdul Aziz Rd., Palatine St., and Almadinah Rd. On the other side, Fig. 2.3b illustrates that the top five mentioned roads/streets names in the tweets about Makkah, which are Makkah-Jeddah highway, Alhaj street, Almadinah Almunawwarah road, Ajyad street, and Alsail road. This result is reasonable where millions of Muslims visited Makkah in Ramadan to perform Umrah and pray in Al-Masjid Al-Haram, which could affect the traffic to/from the city, in addition to the traffic to/from Al-Haram. Ajyad is one
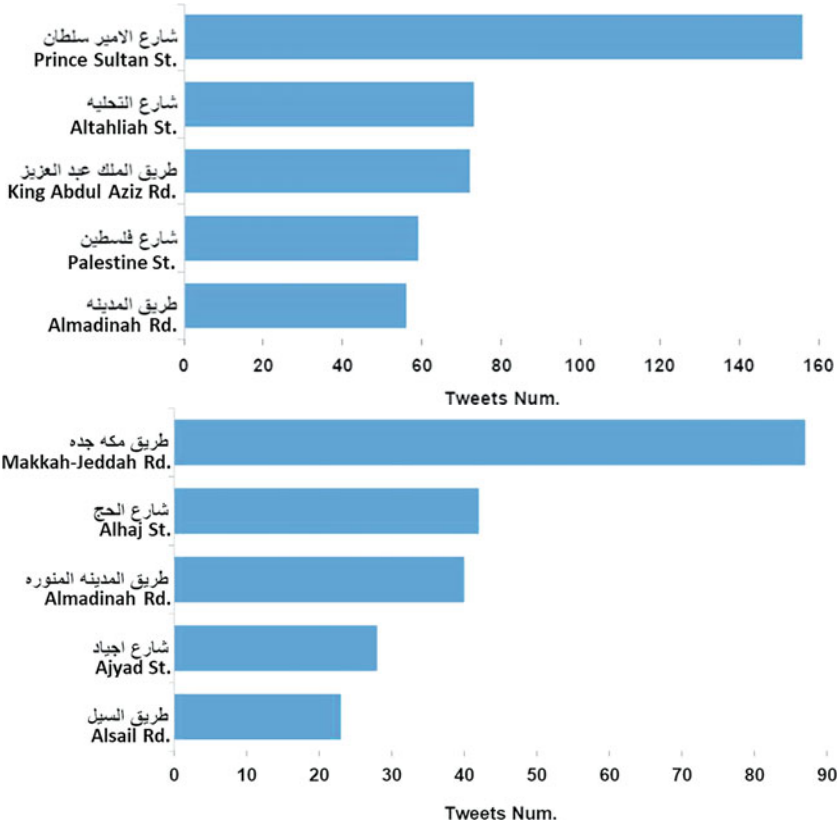
Fig. 2.3 Top mentioned roads/street names. (**a**) Jeddah. (**b**) Makkah

of the main streets leading to Alharam. Alhaj street is one of the main streets in Makkah and connects many districts. The other roads are the main roads connecting Makkah with Jeddah, Al-Madinah, and Al-Ta'if (Alsail Rd.) cities.

Furthermore, Fig. 2.4 illustrates the top detected events in Jeddah and Makkah. The events are detected based on the existing of terms in the created dictionaries. In this work, we exclude the retweets (repost of another user's posts) except when detecting the top mentioned events. The number of retweets is an indication of popularity. Further, it has been implemented to detect events [38]. So, we included the retweets number when detecting the top events.

As shown in Fig. 2.4a, the top three detected events in Jeddah are accidents, fires, and inauguration. To validate our event detection mechanism, we searched in newspapers websites (Okaz, Sabq, etc.) to compare the results. We found that there was a fire in "Extra Store" (on May 28) near Altahliah St., and another building
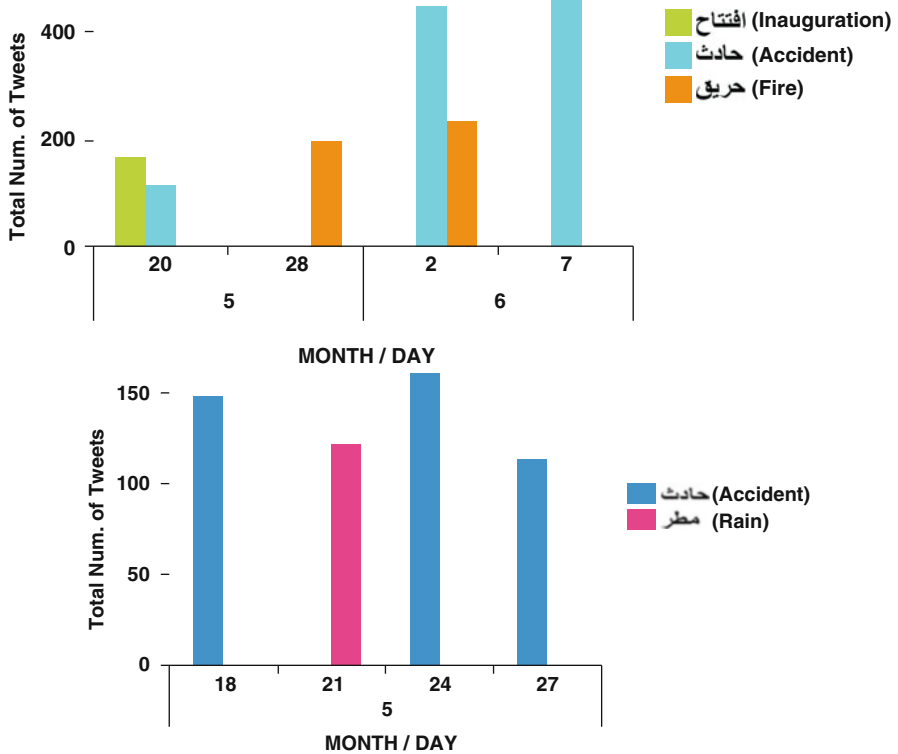
Fig. 2.4 Top detected events. (**a**) Jeddah. (**b**) Makkah

fire (on June 2) leads to 14 injured. In addition, Jeddah Municipality announced that construction work in Al-Andalus Tunnel was done and the tunnel inaugurated at the beginning of Ramadan. We also found articles about several car accidents occurred during Ramadan, one of them was on June 2, where a driver lost control of his car. Further, our tool detected accident on June 7. However, we discovered from searching that the accident occurred on June 5, but there were new posts about it two days later when a nurse honored by the ministry of health for helping injured people in that accident while she is out of work hours.

Moreover, Fig. 2.4b shows the top detected events in Makkah which are rains and accident. We found posts on online newspapers about rains in Makkah on May 21. Additionally, our tool detected several accidents during Ramadan. One of them was on May18. We found details in newspapers articles where there were 9 deaths and 18 injured in a bus accident. Additionally, we found posts about another car accidents (on May 24) in the road connects between Makkah and Al-Madinah cities. From the above discussion, we can notice that the developed tool can automatically detect the traffic events from twitter posts.

Table 2.1 Examples of sentiment classification for driver's feeling and opinions

	Tweets	English Translation	Classification
1	من هذا الشارع بالذات الزحام غير طبيعي	From this street, the congestion is unusual	Negative
2	امطار مكه المكرمة صورة احد شوارع العاصمة المقدسة روعه	Rains in Makkah Almukarramah, a photo of one of the streets in the holy capital, magnificence.	Positive
3	لا يعقل مشروع نفق طريق الأندلس تقاطع شارع فلسطين قبل يومين افتتاحه واليوم صيانه ومسار واحد فقط مما سبب زحمة .. للأسف...	Unbelievable, the tunnel project of Al-Andalus Rd. Palestine street intersection opened two days ago, and today maintenance, only one-way opened, which caused a traffic jam ... Unfortunately ...	Strong negative
4	نفق او كبري او طريق jedgovsa@ الاندلس بجدة ممتاز وصرحة ابداع الي افتتحوه قبل كم يوم او با الكثير شهر	@jedgovsa Alandalus tunnel or bridge or road in Jeddah is excellent and creative, opened a few days or a month ago.	Strong positive

Table 2.1 shows examples of sentiment classification for driver's feelings and opinions. We gave an English translation for non-Arab readers. We provided a literal translation to avoid giving meaning from our side. The tweets are classified into one of 4 sentiment classes based on the total score that is calculated after dividing the text into tokens and identifying the class for each token. For instance, the combination of the two negative terms "congestion" and "unusual" in tweet#1 leads to classifying the tweet as negative. Furthermore, the word "Rain" is labeled as negative where it almost causes negative effect on traffic. However, the existence of the word "magnificence" in tweet#2, which is a strong positive keyword leads to classifying the tweet as positive.

Furthermore, we draw a chart to illustrate the list of the top mentioned words related to the causes of congestions. Figure 2.5 indicates that the word "حادث" (accident) was the most traffic cause mentioned in the collected tweets about traffic in Makkah and Jeddah. Figure 2.6 shows the word cloud for the top used terms about roads and traffic which include street "شارع," road "طريق," accident "حادث," and congestion "زحام."

2.5   Conclusions

In this work, we analyzed Saudi dialect tweets about road traffic conditions. We collected tweets during Ramadan and focused on two large cities (Jeddah and Makkah). We developed our method on SAP HANA, which is an in-memory processing platform to store and analyze the data. The default analysis configuration in SAP HANA is not efficient for Arabic text analysis. So, we created a new configuration file. We added new dictionaries for the Arabic and Saudi dialect
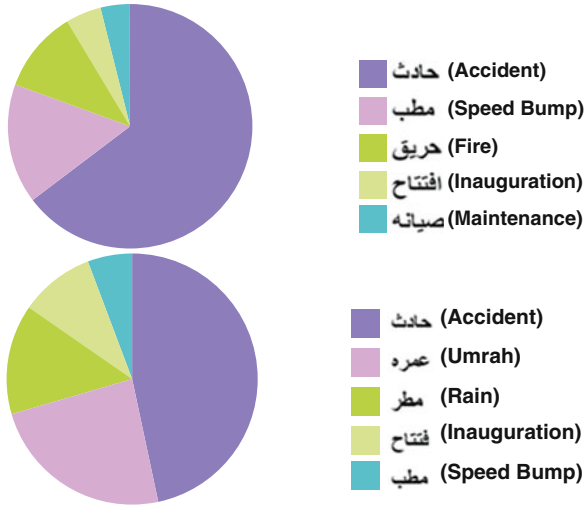
Fig. 2.5 Top mentioned terms related to congestion causes. (**a**) Jeddah. (**b**) Makkah
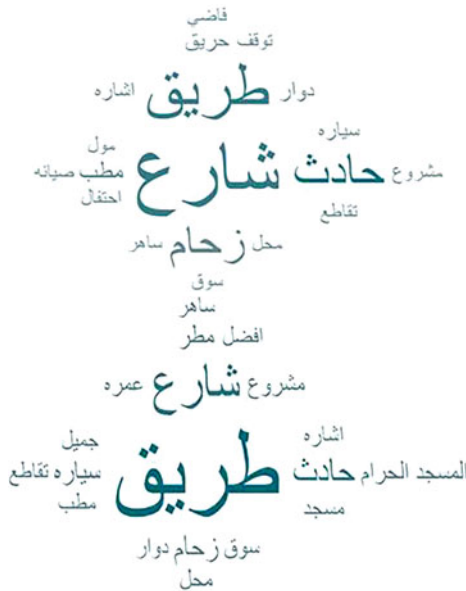

Fig. 2.6 The most frequent terms. (**a**) Jeddah. (**b**) Makkah

keywords related to sentiment, traffic events, and streets names. These dictionaries help in improving tokenization, normalization, and entity extraction. The main contributions of this work are detecting traffic-related events and applying sentiment analysis based on lexicon approach to classify driver's feeling and emotions.

Moreover, we have used SAP Lumira to visualize the results by creating charts. We drew a chart to represent the top mentioned traffic events in the tweets. Additionally, we showed the most frequently mentioned terms related to congestion causes. To validate the proposed event detecting mechanism, we compared the results with data from local newspapers websites. In the future, we plan to measure the accuracy of our proposed sentiment classification approach. Additionally, we will expand our sentiment lexicon and include more words.

References

1. Wang, S., He, L., Stenneth, L., Yu, P.S., Li, Z.: Citywide Traffic Congestion Estimation with Social Media
2. Aldayel, H.K., Azmi, A.M.: Arabic tweets sentiment analysis – a hybrid scheme. J. Inf. Sci. **42**(6), 782–797 (2016)
3. Mourtada, R., Salem, F., Al-Shaer, S.: Citizen engagement and public services in the Arab world: the potential of social media. *Arab Soc. Media Rep.*, no. 2014
4. www.statista.com, Twitter: number of active users 2010–2016 | Statista. 2016
5. Alomari, E., Mehmood, R.: Analysis of Tweets in Arabic Language for Detection of Road Traffic Conditions, pp. 98–110. Springer, Cham (2018)
6. Mehmood, R., Nekovee, M.: Vehicular AD HOC and grid networks: discussion, design and evaluation. In: 14th World Congress on Intelligent Transport Systems, ITS 2007, vol. 2, pp. 1555–1562 (2007)
7. Kanungo, A., Sharma, A., Singla, C.: Smart traffic lights switching and traffic density calculation using video processing. In: 2014 Recent Advances in Engineering and Computational Sciences (RAECS), pp. 1–6 (2014)
8. Wei, L., Dai, H.-Y.: Real-time road congestion detection based on image texture analysis. Procedia Eng. **137**, 196–201 (2016)
9. Gillani, S., Shahzad, F., Qayyum, A., Mehmood, R.: A survey on security in vehicular ad hoc networks, vol. 7865 LNCS. (2013)
10. Alvi, A., Nabi, Z., Greaves, D.J., Mehmood, R.: Intra-vehicular verification and control: a two-pronged approach. Int. J. Veh. Inf. Commun. Syst. **2**(3–4), 248–268 (2011)
11. Alazawi, Z., Altowaijri, S., Mehmood, R., Abdljabar, M.B.: Intelligent disaster management system based on cloud-enabled vehicular networks, in 2011 11th International Conference on ITS Telecommunications, ITST 2011, pp. 361–368 (2011)
12. Alazawi, Z., Abdljabar, M.B., Altowaijri, S., Vegni, A.M., Mehmood, R.: ICDMS: An intelligent cloud based disaster management system for vehicular networks, vol. 7266. Springer, Vilnius, Lithuania (2012)
13. Alazawi, Z., Alani, O., Abdljabar, M.B., Altowaijri, S., Mehmood, R.: A smart disaster management system for future cities, WiMobCity '14. Int. Work. Wirel. Mob. Technol. Smart Cities, pp. 1–10, (2014)
14. Ayres, G., Mehmood, R.: On discovering road traffic information using virtual reality simulations, in 11th International Conference on Computer Modelling and Simulation, UKSim 2009, pp. 411–416 (2009)
15. Mehmood, R., Lu, J.A.: Computational Markovian analysis of large systems. J. Manuf. Technol. Manag. **22**(6), 804–817 (2011)
16. Büscher, M., Coulton, P., Efstratiou, C., Gellersen, H., Hemment, D., Mehmood, R., Sangiorgi, D.: Intelligent mobility systems: some socio-technical challenges and opportunities. In: Mehmood, R., Cerqueira, E., Piesiewicz, R., Chlamtac, I. (eds.) Communications Infrastructure. Systems and Applications in Europe, pp. 140–152. Springer, Berlin (2009)

17. Mehmood, R., Meriton, R., Graham, G., Hennelly, P., Kumar, M.: Exploring the influence of big data on city transport operations: a Markovian approach. Int. J. Oper. Prod. Manag. **37**(1), 75–104 (Jan. 2017)
18. Mehmood, R., Graham, G.: Big data logistics: a health-care transport capacity sharing model. Procedia Comput. Sci. **64**, 1107–1114 (2015)
19. Graham, G., Mehmood, R., Coles, E.: Exploring future cityscapes through urban logistics prototyping: a technical viewpoint. Supply Chain Manag. **20**(3), 341–352 (2015)
20. Schlingensiepen, J., Mehmood, R., Nemtanu, F.C.: Framework for an autonomic transport system in smart cities. Cybern. Inf. Technol. **15**(5), 50–62 (2015)
21. Schlingensiepen, J., Mehmood, R., Nemtanu, F.C., Niculescu, M.: Increasing sustainability of road transport in European Cities and metropolitan areas by Facilitating Autonomic Road Transport Systems (ARTS). In Sustainable Automotive Technologies 2013 Proceedings of the 5th International Conference ICSAT 2013, pp. 201–210 (2014)
22. Schlingensiepen, J., Nemtanu, F.: Autonomic transport management systems—enabler for smart cities, personalized medicine, participation and industry grid/industry 4.0. In: Sladkowski, A., Pamula, W. (eds.) Intelligent Transportation Systems – Problems and Perspectives, pp. 3–35. Springer International Publishing, London (2016)
23. H.A., Alomar, A., Alrashed, N., Alturaiki, I.: How visual analytics unlock insights into traffic incidents in urban areas. In: Business (2017)
24. Mehmood, A., Khan, I.Q., Mir, M.U., Moin, A., Jooma, R.: Vulnerable road users are at greater risk during ramadan—results from road traffic surveillance data. J. Pak. Med. Assoc. **65**(3), 287–291 (2015)
25. D. A. Kurniawan, S. Wibirama, and N. A. Setiawan, Real-time traffic classification with Twitter data mining, 2016
26. D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F.: Real-time detection of traffic from twitter stream analysis. IEEE Trans. Intell. Transp. Syst. **16**(4), 2269–2283 (2015)
27. Ribeiro, S.S., Davis, C.A., Oliveira, D.R.R., Meira, W., Gonçalves, T.S., Pappa, G.L.: Traffic observatory: a system to detect and locate traffic events and conditions using Twitter Sílvio. Proc. 5th Int. Work. Locat. Soc. Networks—LBSN '12, p. 5, (2012)
28. Wongcharoen, S., Senivongse, T.: Twitter analysis of road traffic congestion severity estimation. In 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. (2016)
29. Hanifah, R., Supangkat, S.H., Purwarianti, A.: Twitter information extraction for smart city. In Proc.—2014 Int. Conf. ICT Smart Soc. Smart Syst. Platf. Dev. City Soc. GoeSmart 2014, ICISS 2014, pp. 295–299, (2014)
30. Gu, Y., (Sean) Qian, Z., Chen, F.: From twitter to detector: real-time traffic incident detection using social media data. Transp. Res. Part C Emerg. Technol. **67**, 321–342 (2016)
31. Alifi, M.R., Supangkat, S.H.: Information extraction for traffic congestion in social network. In International Conference on ICT For Smart Society, pp. 20–21 (2016)
32. Suma, S., Mehmood, R., Albugami, N., Katib, I., Albeshri, A.: Enabling next generation logistics and planning for smarter societies. Procedia—Procedia Comput. Sci., pp. 1–6. (2017)
33. Suma, S., Mehmood, R., Albeshri, A.: Automatic event detection in smart cities using big data analytics. In International Conference on Smart Cities, Infrastructure, Technologies and Applications SCITA 2017: Smart societies, Infrastructure, Technologies and Applications, pp. 111–122 (2018)
34. AL-Smadi, M., Qawasmeh, O.: Knowledge-based approach for event extraction from Arabic Tweets. Int. J. Adv. Comput. Sci. Appl. **7**(6), (2016)
35. Hasanain, M., Suwaileh, R., Kutlu, T.M., Elsayed, H.A.: EveTAR: building a large-scale multi-task test collection over Arabic Tweets, arXiv Prepr. arXiv1708.05517., (2017)
36. Alabbas, W., Haider, M., Mansour, A., Epiphaniou, G., Frommholz, I.: Classification of colloquial Arabic Tweets in real- time to detect high-risk floods. Soc. Media, Wearable Web Anal. (Social Media), 2017 Int. Conf. IEEE., 2017
37. Aliane, H., Information, T., Guendouzi, A., Mokrani, A.: Annotating events, time and place expressions in Arabic texts. In Proceedings of Recent Advances in Natural Language Processing, pp. 25–31. (2013)

38. Alsaedi, N., Burnap, P.: Arabic event detection in social media. In LNCS, vol. 9041, pp. 384–401 (2015)
39. Alsaedi, N., Burnap, P., Rana, O.: Can we predict a riot? Disruptive event detection using Twitter, vol. 17, no. 2, (2017)
40. Siddiqui, S., Monem, A.A., Shaalan, K.: Towards improving sentiment analysis in Arabic. In Advances in Intelligent Systems and Computing, vol. 533, pp. 114–123 (2017)
41. Duwairi, S.R.R.M., Marji, R., Sha'ban, N.: Sentiment analysis in Arabic Tweets. In Information and communication systems (icics), 2014 5th international conference on. IEEE, vol. 12, no. 11 (2014)
42. Duwairi, R.M.: Sentiment analysis for dialectical Arabic. In 2015 6th International Conference on Information and Communication Systems, ICICS 2015, pp. 166–170 (2015)
43. Abdul-Mageed, M., Diab, M., Kübler, S.: SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. **28**(1), (2014)
44. Rafea, A., Shoukry, A., Rafea, A.: Sentence-Level Arabic sentiment analysis sentence-level Arabic sentiment analysis. In Collaboration Technologies and Systems (CTS), 2012 International Conference on. IEEE, (2012)
45. Alomari, K.M., Elsherif, H.M., Shaalan, K.: Arabic Tweets sentimental analysis using machine learning. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 602–610 (2017)
46. Abdulla, N.A., Ahmed, N.A., Shehab, M.A., Al-Ayyoub, M., Al-Kabi, M.N., Al-rifai, S.: Towards improving the lexicon-based approach for Arabic sentiment analysis. Int. J. Inf. Technol. Web Eng. **9**(3), 55–71 (2014)
47. Al-Horaibi, L., Khan, M.B.: Sentiment analysis of Arabic Tweets using semantic resources. Int. J. Comput. Inf. Sci. **12**(2), (2016)
48. Al-Hussaini, H., Al-Dossari, H.: A Lexicon-based approach to build service provider reputation from Arabic Tweets in Twitter, (IJACSA). Int. J. Adv. Comput. Sci. Appl. **8**(4), (2017)
49. Al-twairesh, N., Al-khalifa, H., Al-salman, A., Al-ohali, Y.: AraSenTi-tweet: a Corpus for Arabic sentiment analysis of Saudi tweets. Procedia Comput. Sci. **117**, 63–72 (2017)
50. SAP, What is SAP HANA | In Memory Computing and Real Time Analytics, 2016.
51. SAP HANA Web-Based Development Workbench - Introduction to SAP HANA Development - SAP Library.
52. SAP HANA Text Analysis Language Reference Guide, 2016
53. SAP HANA Text Analysis Developer Guide, 2016