

# ATAM: Arabic Traffic Analysis Model for Twitter

Amani AlFarasani<sup>1</sup>, Tahani AlHarthi<sup>2</sup>, Sarah AlHumoud<sup>3</sup>

Computer Science Department, Al-Imam Mohammad Ibn Saud Islamic University (IMSIU)  
Riyadh, Saudi

**Abstract**—Harvesting Twitter for insight and meaning in what is called sentiment analysis (SA) is a major trend stemming from computational linguistics and AI. Industry and academia are interested in maximizing efficiency while mining text to attain the most currently available data and crowdsourcing opinions. In this study, we present the ATAM model for traffic analysis using the data available on Twitter. The model comprises five components that start with data streaming and collection and ends with the road incident prediction through classification. The classification of data is done using a lexicon-based method. The predicted classes are as follows: safe, needs attention, dangerous, and neutral. The data were collected for three months in the city of Riyadh, Saudi Arabia. The model was applied on 10k tweets with an overall accuracy of the model classifying all four classes of 82%.

**Keywords**—Data mining; machine learning; sentiment analysis; unsupervised learning; lexicon-based; support vector machines

## I. INTRODUCTION

The Kingdom of Saudi Arabia is one of the Gulf Cooperation Council (GCC) countries. It is divided into 13 regions; each region is divided into a number of governorates. Riyadh is the capital city of Saudi Arabia with a measured area of 1,554 km<sup>2</sup> [1] and a population of 6,505,509 [2].

According to statistics from the General Authority for Statistics in Kingdom of Saudi Arabia (GASTAT)<sup>1</sup>, 242,851 driving licenses were issued in the Riyadh region in 2016. A total of 141,736 accidents occurred in the same year. Road conditions are to blame for many of the accidents in Riyadh city. Traffic jams, potholes, extreme weather conditions, gas explosions, and malfunctioning traffic lights contribute to the accidents. It is important for travelers to learn about these conditions before making a trip, to improve safety and driving efficiency. According to Internet World Stats statistics, the number of Internet users is over 4 billion, with more than 219 million of them being Arabic users. Also, in 2017, users who provided Arabic web content scored the fourth highest among all Internet users after English, Chinese, and Spanish. In 2018, more than 70% of the Saudi population used the Internet. That number is projected to increase.

Among all social media platforms, Twitter has relatively heavy usage in Saudi Arabia for expressing opinions, advertising, sharing photos, locating information, and discussing various topics. The number of active Twitter users in 2019 exceeded 50% of the total population<sup>2</sup>. Harvesting

available data from Twitter to gain insight and transportation intelligence is a low-cost complementary solution to the infrastructure-based high-cost solutions [3], [4]. Gathering crowdsource opinion data from Twitter using the well-known methodology of Sentiment Analysis (SA) is cheaper and faster than other methods and covers a large number of users in real-time. It is better than using surveys and sensors in terms of cost and timeliness. SA applies supervised or unsupervised machine learning approaches and is implemented using one of the leading programming languages, such as Python or R, or other tools and environments, such as Orange and WEKA [5], [6].

With the high use of Twitter in the region and the raising concerns about road safety in the fast-growing city of Riyadh, Saudi Arabia, the ATAM project aims to provide a model for harvesting road conditions from Twitter. In addition, the lexicons needed to classify data are available upon request for reproducibility. The harvested data are analyzed using SA approaches to provide an instant glimpse of road conditions for drivers and road users. It could also help in instantaneously notifying authorities about road conditions that could result from weather damage or accidents that affect safety. The ATAM model consists of five components: data collection, data preparation, spam filtering, data annotation, and classification. In the next section, we provide background information with related works on SA in general, in the Arabic context, and in terms of traffic analysis. Then, in Section 3, we present the methodology. Section 4 presents the results and discussion. The conclusion of the work is then provided.

## II. BACKGROUND AND RELATED WORK

Utilizing Twitter data to get instantaneous insights into traffic patterns could be done using SA techniques and methods.

### A. Sentiment Analysis

SA is natural language processing (NLP) methodology used to analyze human opinions, emotions, attitudes, and sentiments. Its prominent growth has accompanied the high use of social media, providing vast amounts of data that are available in the public domain.

SA can be implemented using several approaches: Medhat et al. [7] explored these approaches and divided them into two main types. The first is based on machine learning, and the second is a lexicon-based approach. The lexicon-based approach is further divided into the dictionary-based and corpus-based approaches. A corpus-based approach can be statistical or semantic. The machine-learning approach is also divided into supervised learning and unsupervised learning. In the literature, [8, 9, 10, 11, 12] Arabic SA has been accomplished by the following three main steps:

<sup>1</sup> "The General Authority for Statistics": <https://www.stats.gov.sa/en>

<sup>2</sup> "Statista": <https://www.statista.com/statistics/284451/saudi-arabia-social-network-penetration/>

1) *Preprocessing*: After collecting data and before data classification, data need to be cleaned to remove noise. Preprocessing can be done by the following steps: First is data tokenization, which is an essential factor in understanding and manipulating data. This process aids in removing unrelated data, such as usernames and URLs. It also aids in fixing spelling and removing mistakenly repeated characters. This process also works to remove stop words or words that have no polarity significance in the sentences. Second is normalization; and it is a process to reduce the characters in each word to the minimum representable form and to modify multi-form words to a unified form. The former is done by removing diacritics, for example, and the latter is done by removing and replacing the characters that have more than one form to one of its forms [13].

2) *Features extraction*: This process comes after preprocessing, tokenization, and normalization. It mines features from data that can be used to categorize data entries to a class depending on its features [14]. Three types of features are mined [15]. First, morphological features and this consists of semantic, syntactic, or lexico-structural items. The second type is the frequent product feature, which is also referred to as hot features [16]. The third type is implicit features, which are not directly apparent.

3) *Classification*: Classification is the process of dividing and classifying data into two or more classes to facilitate and automate the understanding of data. In SA, data are mainly classified into positive and negative types. Data classification can be done, as stated earlier, using two approaches. First is the supervised learning approach, where data are divided into training data and testing data. Training data are annotated by a human expert labeling the data, which are provided to teach and train the classifier. Testing data are new data used to test and evaluate the performance of the classifier for accuracy. Second, is the unsupervised approach or the rule-based approach where data are classified according to the knowledge provided in labeled lexicons [17].

4) A plethora of researchers are interested in SA for publicly available data streams of social media, as summarized in Table I. Zhou et al. [18] collected a dataset containing 57,000 tweets in the form of 1,000 tweets split into 57 files. The data collection was done in two weeks on the topic of the Australian federal election in 2010. They distributed the data according to sentiment into three

categories: positive, negative, and neutral. Instead of using the “bag-of-words” traditional method, they opted to extract data that include sentiment or words that express a subjective opinion. Measured by category, 65.1% were positive tweets, 77.2% were negative, and 46.2% neutral. The authors were able to identify words with opinions using a rule-based approach with Wilson opinion lexicon [19]. They also measured the strength intensity of positive or negative opinions. They accomplished that using three modules: a feature selection module, which extracts the opinionated words from each sentence; a sentiment identification module, which associates expressed opinions with a relevant entity at each sentence level; and a sentiment aggregation and scoring module, which calculates the sentiment scores for each entity. The sentence intensity was divided into five classes: strong negative (SN), negative (N), neutral (Neu), positive (P), and strong positive (SP). The researchers used a primarily straightforward approach to SA; however, this approach needs to overcome some limitations in certain areas for it to reach its full potential. Some of these areas include distinguishing between parts of speech, taking emotion analysis into account, and utilizing more accurate entity recognition techniques. The authors claimed, “The TSAM model will yield much more accurate results with the above works implemented.” Although they presented the model and plotted the results, they did not provide accuracy data to validate and show the significance of the model.

An important part of SA is the readiness of the data collected, are which has a direct effect on the performance of the classifiers. This is described by Gokulakrishnan et al. [8], who focused on the preprocessing stage of SA, which includes the following steps: replacing emoticons, identifying uppercasing and lowercasing, extracting the URL, determining the punctuation, removing stop words and query terms, compressing words, and removing skewness in the dataset. The datasets collected included 17,000 tweets. The authors used three classifiers: neutral, polar, and irrelevant. They implemented more than one type of algorithm to classify the datasets and to compare them with each other’s performances. They noted that the sequential mining optimization algorithm (SMO) [20] had the best accuracy, where the positive measured 65.1%, the negative was 77.2%, and the neutral was 46.2%. When using the Synthetic Minority Oversampling Technique (SMOTE) [21], the average accuracy of SMO increased from 77.2% to 81.9%.

TABLE I. SENTIMENT ANALYSIS HIGHLIGHT IN ENGLISH LITERATURE

Paper	Context	Classifier	Data Size	Labels	accuracy
[18]	Australia	Lexicon-based SA	57,000	SN (Strong negative), N (Negative), Neu (Neutral), P (Positive) and SP (Strong Positive)	NA
[8]	Sri Lanka	Naïve Bayes, Random Forest, Support Vector Machines and Sequential mining optimization	17,000	Neutral, Polar (Positive/Negative) and Irrelevant	SMO 77.2%
[9]	South Korea	Knowledge generator, knowledge enhancer, and synonym binder	40,000	Positive, Negative and Neutral	accuracy was 55%.

Another work on precise classification was done by Batool et al. [9], proposed using Archivist, which is a service that uses Twitter API to find and archived tweets. Then they used Alchemy API, which utilizes NLP and machine-learning algorithms to analyze content. They collected a dataset of 40,000 tweets of different categories for testing and verification in 43 days. Then they divided these data depending on sentiment into three categories: positive, negative, and neutral. They used a knowledge-enhancer module, which adds additional knowledge that was not extracted as keywords by Alchemy API. Their accuracy score was 55%.

### B. Arabic Sentiment Analysis

The Arabic language could be divided into three types [11]. First is classical Arabic, which is used in the Quran holy book and prayers. Second, is Modern Standard Arabic (MSA), which is used in formal contexts, such as in books, education, and news. The third is the Dialectal Arabic (DA), which is used informally in verbal communications and is used recently in written communication with the use of social media and short messages. These forms of the language result in lexical, morphological, and grammatical differences resulting in the difficulty of developing one Arabic NLP application to process data from the different varieties [22].

Besides, Arabic NLP applications face the challenge of encoding, which is the representation of the language symbols in computers, especially when representing the different shapes of the same letter or the diacritics. Unicode is the actual current standard for encoding a large number of language symbols including Arabic, such as the Arabic letter ك (U+0643) and the Persian ك (U+06A9) using the same shape ك, which adds confusion when the Arabic letter is written using a Persian keyboard [10] [23] [24] [11].

Another challenge to Arabic SA is the lack of gold-standard corpora, quality resources, accurate stemmers, and tools compared to English. For that, the research in Arabic NLP is still in its early stages, needing more resources and

efforts. This paper aims to provide a model to classify Arabic text in the traffic domain, contributing by enriching this field with the ATAM model.

In the following, we present the highlights of previous studies on Arabic SA, which are summarized in Table II. Ibrahim et al. [25] used ArSeLEX Lexicon with a collection of 5244 words. First, they used an AMIRA Part of Speech (POS) tag [26] to extract the words with a higher likelihood to be sentimental, such as adjectives, nouns, and verbs. Second, they removed redundant words. Then, each of the remaining words is translated, and all its synonyms are fetched. The output of the dataset was 300 positive, 2,829 negative, and 412 neutral terms. The Arabic variety was MSA and Arabic Egyptian dialect. The highest accuracy they reached using the SVM classifier measured 95%.

Khasawneh et al. [27] collected 1,500 Arabic comments and audio segments from the Twitter website. Then, the data were broken down according to news type into sports or economy. This was done using MSA analysis by manually constructing 13 dictionaries, 6 of which were for positive and negative Arabic text, 2 for audio files, 3 for positive or negative or neutral symbols, and 2 for special characters. The results were evaluated by using two machine-learning classifier techniques: the bagging and Boosting techniques. The bagging technique accuracy result was 82.95%, while the boosting technique's accuracy score was 64.52%.

Albraheem et al. [28] used the NODEXL tool to retrieve tweets and compiled 100 tweets with Saudi hashtags. This was a small number that could hardly be reliable enough to draw any learning conclusions. The number of positive words that their model found in the tweets was 33 while the number of positive words detected by human language experts was 40. The accuracy of positive words was 82.5%, while the accuracy of negative words was 71.01%. The accuracy of all tweets was 73%.

TABLE II. SENTIMENT ANALYSIS HIGHLIGHTS IN ARABIC LITERATURE

Paper	Lexicon	Tools/ Languages	Size	Sentiment Score	Arabic variety
[25]	ArSeLEX	AMIRA	5,244 word	Positive 300 Negative 2,829 Neutral 412	MSA & Egyptian dialectal Arabic
[27]	NA	NA	1,500 Arabic comments	Sports: Positive 369 Negative 171 News: Positive 83 Negative 316 Economics: Positive 234 Negative 327	MSA & dialectal Arabic
[28]	NA	NODEXL tool	100 Arabic tweets extracted from some of the Saudi hashtags	The positive words were 82.5% and the negative words were 71.01%	Dialectal Arabic
[29]	created by authors	R	3,000 tweets	The accuracy of the unsupervised approach was 81.70%.	MSA & dialectal Arabic

Alhumoud et al. [29] implemented a hybrid learning approach that combines lexicon and supervised approaches compared to the supervised and unsupervised learning approaches. Both the supervised classifier and the hybrid classifier trained on 3,000 tweets collected from three domains. The unsupervised approach has two dictionaries, positive and negative. The training dataset contains 3,690 sentimental words, which are built from rows of single sentimental words and their labels. The training datasets included 1,370 positive words, and 2,320 negative words, 1,000 MSA sentimental words, and 2,690 Saudi dialect sentimental words. The accuracy of the unsupervised approach was 81.70%.

### C. Sentiment Analysis in Traffic

SA in traffic is concerned with tapping into the available datasets on traffic with the aim of inferring meaning, indicators, and safety signals that foster more efficient driving. The following related work highlights the available research on SA in the traffic domain with a Twitter dataset source.

Kurniawan et al. [30] collected data consisting of 110,449 tweets for seven days from official traffic accounts from the Indonesian province Yogyakarta. They used three algorithms for machine-learning, namely Naïve Bayes (NB), a Support Vector Machine (SVM), and a Decision Tree (DT). The data were classified into two categories: traffic and non-traffic tweets. The results show that the SVM provided the best performance, as its classification accuracy in balanced and imbalanced data was 99.77% and 99.87%, respectively.

Andrea et al. [31] aimed to detect real-time traffic accidents from data consisting of 2,649 tweets using n-labelled SUMs and classified according to Status Update Message (SUM), which is the user message shared in social networks and class labels related to traffic events. The highest value reached with the SVM classifier was 95.75 %. One of the drawbacks of this study, which was conducted in Italy, is the lack of a data collection period or a list of the number of words.

Lee et al. [32] collected data from 22,353 Korean messages within three months in 2014. The number of Twitter messages (62,495) were collected from 5,247 users. They collected their data using Traffic Information Producers (TIPs) and Opinion Leaders (OLs) and keyword and network analysis. The data were classified into categories including traffic conditions, locations, and instructions and were measured at 90% accuracy.

Wang et al. [33] collected 245,568 tweets on traffic events in Chicago, USA. The classification was done using the EM algorithm to classify data into three classes: slow traffic, accidents, and other road conditions (e.g., construction). The dataset sizes were 163,742, 77,454 and 4,372, respectively, and the accuracy value was 85%.

Alhumoud [34] presents a framework for Arabic Twitter content analysis to gain traffic insight applied in the city of Riyadh, Saudi Arabia. The study was done with a dataset of more than 1 million tweets collected within three months. The proposed model comprised three main components: data acquisition, data analysis, and a reverse geotagging scheme (RGS). The data acquisition phase utilized AsterixDB to collect tweets and perform preliminary preprocessing. AsterixDB is a “highly scalable data management system that can store, index, and manage semi-structured data.” In the data analysis phase, the data were analyzed using the hazard classifier based on the transportation hazard index (THI), which is a lexicon provided by the author yielding one of four possible hazard intensities for each tweet. The hazards were classified into four types: accident incidents, weather incidents, negative road incidents, and positive road incidents. The results showed that 13% of the dataset reported traffic-related incidents with an overall precision of 55% and 87% for incidents identification prediction without and with reverse geotagging, respectively.

A summary of SA studies on traffic is depicted in Table III.

TABLE III. SENTIMENT ANALYSIS IN TRAFFIC

Paper	Data Size	Collection	Classification	Label	Accuracy
[30]	NA	Using Twitter Streaming API	NB & SVM and DT	Traffic tweets and non-traffic tweets	balanced and imbalanced data were 99.77% and 99.87%
[31]	NA	Using Twitter Streaming API	SVM	NA	SVM classifier 95.75 %
[32]	22,353	(API) & using a keyword analysis and a network analysis.	TIPs & OLs	NA	90%.
[33]	245,568	API	Sequential importance sampling-based EM algorithm	Tweets on slow traffic, accidents and other road conditions.	85 %.
[34]	1M	AsterixDB	Lexicon based using THI	Incidents: accident, weather, negative, positive	55% and 87%

### III. METHODOLOGY

The ATAM system comprises five components. Those components include data collection using two methods to be explained in the next section. The second component is data preprocessing and denoising with state of the art text preprocessing techniques and normalization. The third component is the spam filtering according to the rules implemented in [35]. The fourth component is an annotation; that is, labeling the corpus by human experts into the desired four classes to train the model for correct classification. The fifth and final component is the classifier, that classifies data into four classes using a rule-based classifier. The ATAM system components are depicted in Fig. 1. Following is a more detailed explanation of the system.

#### A. Collecting Data

Twitter data were collected using R with Twitter API, which allows accessing tweets and collecting them using two approaches. First, a streaming function that allows the collecting of tweets in real time based on the provided street lexicon. The second approach involves collecting specific user tweets using the userTimeline function, which allows researchers to pull the latest 3,200 tweets from a user timeline. The accounts that are scrapped are known for traffic tweets and hazardous road conditions in the city of Riyadh. The number of collected tweets reached 292,965 by using 44 street keywords in a three months period from September 2017 to November 2017. The streets under consideration are in the city of Riyadh, Saudi Arabia.

#### B. Preprocessing

After streaming Twitter data, the data were prepared for analysis. Preprocessing involves nine actions: (1) removing repeated letters (i.e., converting “liiiiife” into “life”), except words that match the road lexicon are kept unchanged; (2) normalizing by converting multiple forms of a letter into one uniform letter [i.e., (ة to ة), (ة to ة), (ي to ي), and (أ/إ to ا)]; (3) removing numbers, spaces, and the tatweel, which is the Arabic letter (-), or (4) (المالك) removing any non-Arabic letters; (5) removing punctuation marks (e.g., “.,;?!), except hashes (#) (which is required in the spam filtering step, that classifies a tweet with more than four hashes as a spam tweet); (6) removing Arabic diacritics, such as َ; (7) removing stop words, which are provided in [36]; (8) removing words that have less than two letters; (9) and stemming.

The stemming process involves the extraction of a word root to enhance the classifier accuracy by merging many word forms into one root form [37]. The Arabic language has a composite morphology structure that makes root extraction more complicated and limits the stemming to removing prefixes and suffixes [38]. However, there are several algorithms can simplify extracting roots. These algorithms follow some rules for removing prefixes and suffixes to produce proper stemming, such as the AlKabi [39], Ghawanmeh [40], Hmeidi [41], Khoja [42] and WSS-Based algorithms [37].

The Light10 stemmer [43], which is claimed to be the best available stemmer, works by solely removing the initial letter

ها, ان, ات, ون, ين, ), and suffix (ال, وال, بال, كال, فال, لل), prefix (و, ي, يه, ية, م, ق, ي), and this may not result in an accurate root extraction. In the case of this research, the arabicStemR package in R developed by Nielsen in MIT was used. However, the stemmer included suffix and prefix elimination, and that changed the meaning of some important keywords, such as street names (e.g., “شارع الستين” was transformed into “شارع ست” after stemming). For this reason and because of the limited added value by stemming in the study’s dataset, the stemming step was ignored.

As for removing the stop words, it was postponed preceding the annotation step to preserve the meaning and clarity of the sentences and to enable correct annotation by the experts.

#### C. Spam Filtering

One of the significant challenges in studying datasets from Twitter is the high volume of noise or spam tweets. Spam data are unrelated data that are collected with the target data, including advertisements and news. As the dataset size was large, the need for an automated spam filtering was inevitable. We used the algorithm provided in [35] where tweets with URLs, phone numbers, more than four hashtags, and duplicated tweets are classified as spam. The algorithm also implements a rule-based classifier with a spam lexicon. Also, in this study, tweets that not related to streets or tweets with less than three words were classified as spam. The number of remaining tweets after spam filtering decreased by 96%, leaving 11,037 tweets.

#### D. Annotation

In this step, the resulting dataset from the previous step undergoes labeling by two expert Arabic speakers. The procedure of annotation was as follows. Using the instructions specified by the authors, the two experts labeled 5,781 data entry items into one of the following labels: neutral, safe, needs attention, and dangerous. By agreement of the two experts on a data entry label, a data entry was accepted with the given label. If they disagreed, the data entry was eliminated. After annotation, the number of safe tweets accounted for 8%, while the dangerous tweets accounted for 6%. Tweets that need attention accounted for 18%, while neutral tweets reached 68%.

#### E. Arabic Traffic Analysis Model for Twitter

The ATAM model implements a rule-based classifier that classifies data into four classes: safe, needs attention, dangerous, and neutral. The classifier utilizes three lexicons built using the gulf region dialect, which is commonly used by Saudi Twitter users. After applying the previous steps, the size of the dataset was 10,175 tweets. Technically, the ATAM model implements four counters that count the occurrences of the four different classes; in each tweet in the dataset by matching each keyword in the lexicon to the available dataset using the R language. Then, each tweet is classified according to the most repeated class label. If the labels from each class occur in one tweet equally, the highest occurring class in severity is assigned with the following priority: dangerous, needs attention, safe, and neutral. The algorithm is shown in Fig. 2.

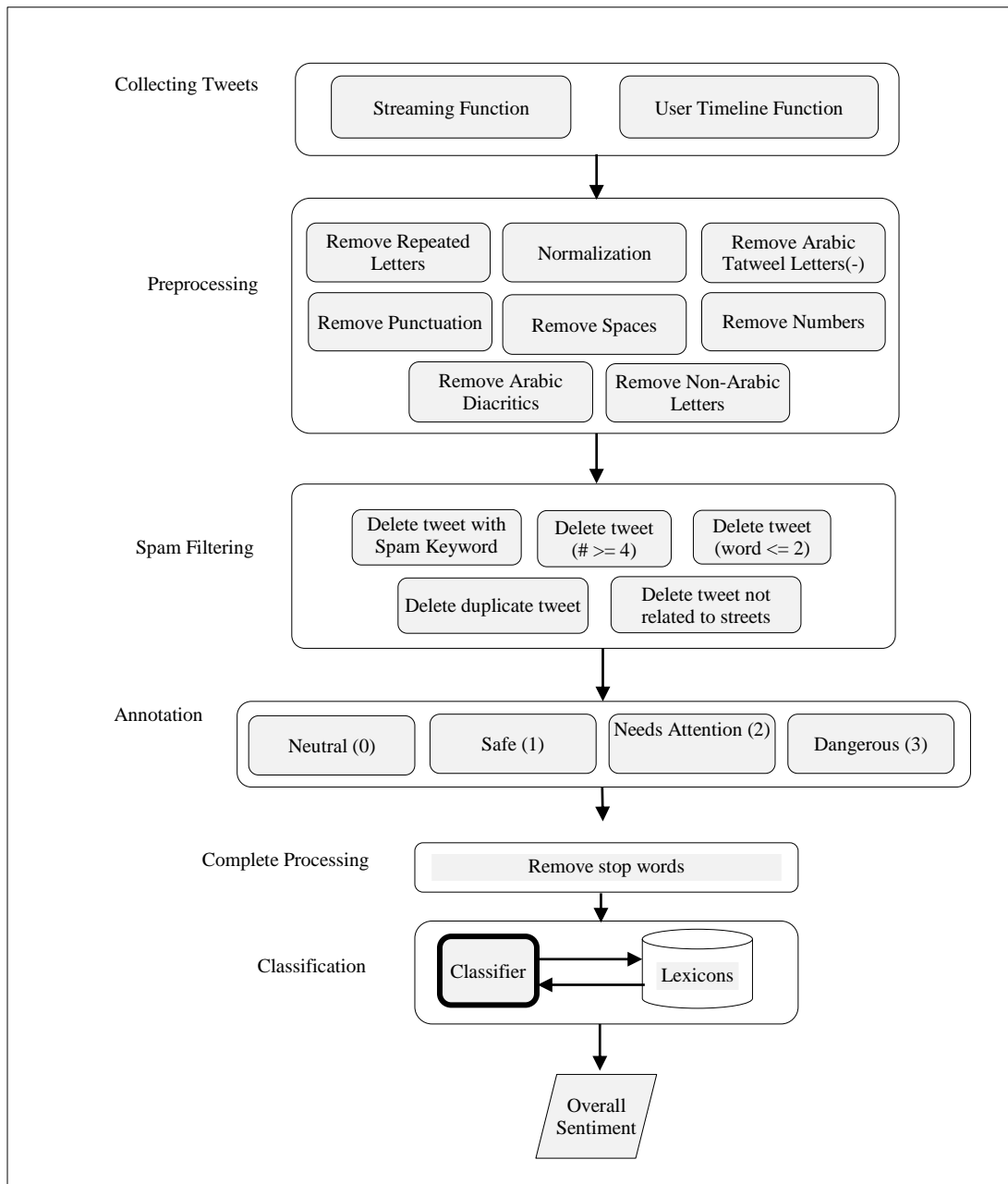


Fig. 1. ATAM System Architecture.

#### IV. RESULTS AND DISCUSSION

After building the ATAM model, we tested and evaluated the accuracy of this model using equation (1), which is one of the most common metrics used to measure performance. Accuracy was measured for all four classes: safe, dangerous, needs attention, and neutral over 300 tweets.

Equation 1 shows the accuracy formula, where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively. True positive stands for test results that detect the condition when the condition is present. True negative is when it does not detect the condition when the

condition is absent. False positive is when it detects a condition when the condition is absent. Finally, false negative denotes when it does not detect the condition when the condition is present.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

The ATAM model was applied to 10,175 tweets, and 300 of them were studied and tested to calculate the accuracy. The neutral data reached 44% of the total tweets, while safe data was 14%, needs attention data was 25%, and dangerous data was 17% of the total tweets. To calculate accuracy for the dataset under consideration, TP, TN, FP, and FN were

calculated for each label. Those values are depicted in Table IV. As the table shows, the maximum accuracy was for the class “needs attention,” with 88%. Then, the second highest was the “dangerous” class, scoring 86% for the accuracy. The “safe” class scored 85%, while the lowest accuracy score was for the “neutral” class, with 70% accuracy. The average

accuracy scored by the ATAM model was 82%. The low score for the neutral class could be explained by the incorrect classification of tweets of significant incident by the model. This could be improved by enlarging the lexicons to include more incident keywords.

#### ATAM Algorithm

**INPUT:** Tweets T, Safe Lexicon SL, Needs attention lexicons NL, Dangerous lexicons DL, Neutral N

**OUTPUT:** L = {Safe, Needs Attention, Dangerous, Neutral}

**INITIALIZATION:** L= 0, Lsafe = 0, Lneedsattention = 0, Ldangerous = 0, where Lsafe: Sum of safe words, Lneedsattention: Sum of the needs attention words, Ldangerous: Sum of the dangerous words, L: to get the highest between Lsafe, Lneedsattention or Lneedsattention

```
1. For each  $T_i \in T$ 
  1. For each  $SL_i \in SL$ 
    1. If  $SL_i \in T_i$  then
      1. Lsafe = Lsafe + 1
    2. End if
  2. End for
  3. For each  $NL_i \in NL$ 
    1. If  $NL_i \in T_i$  then
      1. Lneedsattention = Lneedsattention + 1
    2. End if
  4. End for
  5. For each  $DL_i \in DL$ 
    1. If  $DL_i \in T_i$  then
      1. Ldangerous = Ldangerous + 1
    2. End if
  6. End for
  7. If Lsafe > Lneedsattention and Lsafe > Ldangerous then
    1. L = Safe
  8. End if
  9. If Lneedsattention > Lsafe and Lneedsattention > Ldangerous then
    1. L = Needs Attention
  10. End if
  11. If Ldangerous > Lneedsattention and Ldangerous > Lsafe then
    1. L = Dangerous
  12. End if
  13. If Ldangerous == Lsafe or Ldangerous == Lneedsattention and Ldangerous != 0 then
    1. L = Dangerous
  14. End if
  15. If Lneedsattention == Lsafe and Lneedsattention > Ldangerous and Lneedsattention != 0 then
    1. L = Needs Attention
  16. End if
  17. If Lsafe == 0 and Lneedsattention == 0 and Ldangerous == 0 then
    1. L = Neutral
  18. End if
End For
Return L
```

End ATAM Algorithm

Fig. 2. Arabic Traffic Analysis Model.

TABLE IV. THE ACCURACY OF 300 TWEETS

<b>Actual \ Predicted</b>	<b>Safe</b>	<b>Needs Attention</b>	<b>Dangerous</b>	<b>Neutral</b>	<b>Acc</b>
Safe	24	1	0	0	85%
Needs Attention	3	43	2	1	88%
Dangerous	0	0	28	0	86%
Neutral	48	31	45	205	70%

The comparably high performance of these results could be due to the accuracy and lack of duplication of the keywords that were used in the lexicon dictionaries for each class. The ATAM model ensures that the tweets are appropriately categorized after calculating the number of lexicons' words that relate to a specific label in one tweet. As explained earlier, if more than one class is represented in a tweet, then the final classification result is assigned to the more severe class. For example, the tweet “ طريق عثمان فاضي بس يحتاج إعادة زفلة فهو مليون ” which means “Uthman road is not busy but needs construction; it has a lot of dangerous holes” holds two sentiments, dangerous and safe; therefore, we programmed our model to assign the final classification for this tweet as dangerous.

### V. CONCLUSION

Twitter traffic analysis serves as a timely and complimentary solution to the costlier infrastructure-based sensors and GPS systems. English text analysis enjoys the abundance of gold-standard corpora and resources. However, in Arabic, text analysis is still in an early stage where the resources scarcity and language nature bring huge challenges to the research. This study presents an Arabic Traffic Analysis Model (ATAM) to tackle this area of research. This model aims to mine related Arabic texts from Twitter to present instantaneous pivots on traffic incidents. These incidents fall into four categories: safe, needs attention, dangerous, and neutral. For this study, we collected around 300k tweets in a period of three months. The tweets were subject to spam filtering, leaving a data size of 10,000 related tweets. Additionally, half of those tweets, were annotated by expert Arabic speakers to measure classifier accuracy. The results showed that the overall accuracy was 82% for all four classes. As a future work, we aim to build a web service for live streaming and classifications.

### REFERENCES

[1] S. Alaamer, "Characterization of 137Cs in Riyadh Saudi Arabia," World Journal of Nuclear Science and Technology, vol. 2, no. 4, pp. 161-164, 2012.

[2] A. A. Salam, "Population and Household Census, Kingdom of Saudi Arabia 2010: Facts and," International Journal of Humanities and Social Science, vol. 3, no. 16, 2013.

[3] Y. Gu, Z. Qian and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," Transportation Research Part C: Emerging Technologies, vol. 67, pp. 321-342, 2016.

[4] S. Alhumoud, "Twitter Analysis for Intelligent Transportation," The Computer Journal, vol. 61, no. 12, 2018.

[5] S. Alhumoud, T. Albuhaire and M. Altuwajri, "Arabic sentiment analysis using WEKA a hybrid learning approach," in 7th International

Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Portugal, 2015.

[6] S. O. Alhumoud, M. I. Altuwajri, T. M. Albuhaire and W. M. Alohaideb, "Survey on Arabic Sentiment Analysis in Twitter," International Journal of Computer and Information Engineering, vol. 9, no. 1, 2015.

[7] W. Medhat, A. Hassan and HodaKorashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[8] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera, "Opinion mining and sentiment analysis on a Twitter data stream," in International Conference on Advances in ICT for Emerging Regions (ICTer2012), 2012.

[9] R. Batool, A. M. Khattak, J. Maqbool and S. Le, "Precise tweet classification and sentiment analysis," in IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS), Japan, 2013.

[10] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in International Conference on Collaboration Technologies and Systems (CTS), USA, 2012.

[11] N. Y. Habash, Introduction to Arabic Natural Language Processing, Morgan & Claypool publishers, 2010.

[12] S. Alowaidi, M. Saleh and O. Abulnaja, "Semantic Sentiment Analysis of Arabic Texts," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 8, 2017.

[13] B. Han, P. Cook and T. Baldwin, "Lexical normalization for social media text," Journal ACM Transactions on Intelligent Systems and Technology (TIST), vol. 4, no. 1, p. 5:1-5:27, 2013.

[14] J. Pereira, Handbook of Research on Personal Autonomy Technologies and Disability Informatics, IGI Global, 2010.

[15] M. Z. Asghar, A. Khan, S. Ahmad and F. M. Kundi, "A Review of Feature Extraction in Sentiment Analysis," Journal of Basic and Applied Scientific Research, vol. 4, no. 3, pp. 181-186, 2014.

[16] M. Hu and B. Liu, "Mining opinion features in customer reviews," in AAAI'04 Proceedings of the 19th national conference on Artificial intelligence, California, 2004.

[17] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.

[18] X. Zhou, X. Tao, J. Yong and Z. Yang, "Sentiment analysis on tweets for social events," in IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Canada, 2013.

[19] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceeding HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Canada, 2005.

[20] J. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," Microsoft research technical report, 1998.

[21] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse and A. Napolitano, "Building Useful Models from Imbalanced Data with Sampling and Boosting," in FLAIRS Conference, 2008.

[22] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing, vol. 8, no. 4, p. 22, 2009.

[23] W. Salloum and N. Habash, "ADAM: Analyzer for Dialectal Arabic Morphology," Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 4, pp. 372-378, 2014.

[24] S. AlOtaibi and M. B. Khan, "Sentiment Analysis Challenges of Informal Arabic Language," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 8, no. 2, 2017.

[25] H. S. Ibrahim, S. M. Abdou and M. Gheith, "Sentiment Analysis For Modern Standard Arabic And Colloquial," International Journal on Natural Language Computing (IJNLC), vol. 4, no. 2, 2015.

[26] M. Diab, "Towards an optimal POS tag set for Modern Standard Arabic processing," in Proceedings of Recent Advances in Natural Language Processing (RANLP), Bulgaria, 2007.

[27] R. T. Khasawneh, H. A. Wahsheh, I. M. Alsmadi and M. N. Al-Kabi, "Arabic sentiment polarity identification using a hybrid approach," in



- International Conference on Information and Communication Systems (ICICS), Jordan, 2015.
- [28] L. Albraheem and H. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services, Indonesia, 2012.
- [29] S. Alhumoud, T. Albuhairei and W. Alohaide, "Hybrid sentiment analyser for Arabic tweets using R," in 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Portugal, 2015.
- [30] D. A. Kurniawan, S. Wibirama and N. A. Setiawan, "Real-time traffic classification with Twitter data mining," in 8th International Conference on Information Technology and Electrical Engineering, Indonesia, 2016.
- [31] E. D'Andrea, P. Ducange, B. Lazzarini and F. Marcelloni, "Real-Time Detection of Traffic From Twitter Stream Analysis," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 2269-2283, 2015.
- [32] H. P. Lee, D.-P. Hong, E. Han and S. H. K. Yun, "Analysis of the characteristics of expressway traffic information propagation using Twitter," KSCE Journal of Civil Engineering, vol. 20, no. 7, pp. 2587-2597, 2016.
- [33] S. Wang, F. Li, L. Stenneth and P. S. Yu, "Enhancing Traffic Congestion Estimation with Social Media by Coupled Hidden Markov Model," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2016.
- [34] S. Alhumoud, "Twitter Analysis for Intelligent Transportation," The Computer Journal, vol. 61, no. 12, 2018.
- [35] N. A. Twairesh, M. A. Tuwaijri, A. A. Moammar and S. A. Humoud, "Arabic Spam Detection in Twitter," in Language Resources and Evaluation Conference, Slovenia, 2016.
- [36] A. Alajmi and E. m. Saad, "Toward an ARABIC Stop-Words List Generation," International Journal of Computer Applications, vol. 46, no. 8, p. 6, 2012.
- [37] Q. Yaseen and I. Hmeidi, "Extracting the roots of Arabic words without removing affixes," Journal of Information Science, vol. 40, no. 3, pp. 376-385, 2014.
- [38] A. Alsaad and M. Abbod, "Arabic Text Root Extraction via Morphological Analysis and Linguistic Constraints," in 16th International Conference on Computer Modelling and Simulation, UK, 2014.
- [39] M. N. Al-Kabi and R. S. Al- Mustafa, "Arabic Root Based Stemmer," in International Arab Conference on Information Technology, Jordan, 2006.
- [40] S. Ghwanmeh, G. Kanaan, R. Al-Shalabi and S. Rabab'ah, "Enhanced Algorithm for Extracting the Root of Arabic Words," in Sixth International Conference on Computer Graphics, Imaging and Visualization, China, 2009.
- [41] I. Hmeidi, R. Al-Shalabi and D. A. T. Al-Taani, "A Novel Approach to the Extraction of Roots from Arabic Words Using Bigrams," Journal of the American Society for Information Science and Technology, p. 583-591, 02 12 2009.
- [42] S. Khoja and R. Garside, "Stemming Arabic Text," Computing Department, Lancaster University, 1999.
- [43] L. S. Larkey, L. Ballesteros and M. E. Connell, "Light Stemming for Arabic Information Retrieval," TLTB, vol. 38, pp. 221-243, 2007.