Tb-SAC: Topic-based and Sentiment Classification for Saudi Dialects Tweets

Sara Alzahrani , Fatimah Alruwaili , Dimah Alahmadi and Kawther Saeedi

Faculty of Computer and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

Summary

Recently, sentiment analysis has received a lot of attention from researchers in text mining and data analysis. The studies have significantly expanded to include different languages from several sources that were employed to create a corpus to serve researchers in various shapes, sizes, and purposes. Locally, a lot of effort is spent on analyzing sentiment for Arabic texts, for both Modern Standard Arabic (MSA) and vernacular dialects. However, the researches concerned with creating a corpus based on the topic was relatively few. In this paper, we present Tb-SAC as extracted corpora from Twitter, especially from Saudi dialects. The corpus contains 4301 tweets, which labeled based on sentiments using a three-point scale: positive, negative, and neutral. The corpus classify based on tweet topics into five main topics obtained from analyzing the gold set with 200 tweets. The topics were Personal, Religion, Coronavirus, Entertainment, Other (Education, Economy, Sport, Food, Health, Social Media, Distance Working, Technology, Comedy, and Politics). Then, we performed the annotation process manually, besides applying eleven different classification models and validate the corpus by cross-validation model.

Key words:
Natural language processing (NLP); Sentiment analysis (SA); Topic-based; Saudi Dialects; Twitter; and Annotation.

1. Introduction

Sentiment analysis (SA) is considered as a classification algorithm that extracts the overall sentiment of a text [1]. SA is one of the fields that has attracted attention nowadays and become vital by an extensive range of real-world applications [2]. SA is based on an annotated corpus to train and test the classifier model; the created corpus should be labeled to be readable for the machine [3]. Nevertheless, just adding sentiment analysis to the corpus may not be optimal for researches in some narrow topics [4].

The most research in topic-based sentiment analysis algorithms is for the English language because of the massive amounts of English corpora that support topic-based sentiment analysis tasks. Where the rest of the corpora in other languages such as the Arabic language are still limited [1]. Arabic language is the fourth-ranked between the universal languages and becomes the fastest growing language on the web [3]. Also, Arabic considers as rich and complex a morphologically language, where one lemma may have a lot of forms, which complex the work of SA algorithms [5].

Manuscript received September 5, 2020
Manuscript revised September 20, 2020

DOI: 10.22937/IJCSNS.2020.20.09.6

Huge differences among spoken Arabic country's culture and dialect, one word may express different sentiment [3]. Furthermore, due to the rare in Arabic language corpora that include both topic identification and target of the sentiment in the annotation [1]. Therefore, this paper aims to create a topic-based sentiment analysis corpus in the Arabic language.

The enormous amounts of data available on the Web., especially in social media, as users can write and express whatever they want freely, wither it's personal or just popular subjects. This is providing a great opportunity for the researchers that work on the SA fields [2]. And the most used social media platform in Arabic countries is Twitter, where 11.1 million Twitter users from Arab regain with 27.4 million tweets generated daily [1]. In Saudi Arabia, the Twitter platform is the most used among other social media, where Riyadh city is considered as the 10th most active city on Twitter [3]. And due to the huge differences among the Arabic country's dialect, which mention previously. Accordingly, this paper aims to present Tb-SAC, an Arabic multi topic-based sentiment analysis corpus that has been created from tweets generated by users in Saudi Arabia only, which can be very beneficial for economic, social, and cultural researchers in Saudi Arabia. The corpus mainly objectives can be as follows:

- Adding more resources to Arabic dialect resources.
- Investigating the dialect Saudi topics in twitter and the related sentiment.
- Establishing a clear guideline for developing and annotating Arabic corpus and address challenges in the process.
- Developing several benchmark classification algorithms to predict topics and test the proposed corpus.

This paper includes the five following sections. First, investigate the previous studies on Arabic sentiment and topic-based corpus. In section 2, to illustrate the gap that fills by this paper. Then explain the development of the Tb-SAC corpus process in detail is presented in section 3. Also, the paper provides obstacles and challenges in section 4. Finally, the last section will discuss the conclusion and future work of this paper.

2. Related Works

The term "sentiment analysis" is used to explain the study and analysis of people's sentiments, opinions, emotions, evaluations, and feelings related to specific products or services they have tried. This analysis aims to achieve accurate results

that reflect people's opinions to be used in research for commercial, social, educational, and other purposes. Researchers began searching for opinions and feelings since the beginning of 2000, but the term sentiment analysis first appeared in [6]. What strengthened this field is the emergence of social media, which can be used to collect a huge number of databases for users' opinions [7]. To build and evaluate a sentiment analysis model, a corpus with pre-labeled sentiment must be available to be applied. Corpus is texts that collected and formed in a way to be machine-readable, which drove by the searching in linguistic phenomena. Usually, the corpus is in different types due to the various purpose of their creation [8]. Here we will illustrate the most common corpus of the Arabic language.

A. Arabic corpora in twitter and sentiment analysis

In a study for Zaghouani, he surveyed to investigate the free Arab corpus available. He collected up to 66 corpora and classified them into six main groups, which are: "1- raw text corpora. 2- annotated corpora. 3- lexicons. 4- speech corpora. 5- handwriting recognition corpora. 6- different corpora types." [9]. Al-Sulaiti and Atwell created the first Arabic corpora available for free. These corpora relied on gathering words from news sources, which are newspapers and magazines, with more than a million words [10]. Numerous researches have followed this approach in collecting texts such as "the Open Source Arabic Corpora (OSAC)" [11] and "Akhbar Al Khaleej" [12] and others. After that, many efforts continued in establishing a corpus concerned with the Modern Standard Arabic (MSA), one of them was presented in AWATIF which was a labeled corpus for subjectivity and sentiment analysis (SSA) [13]. In another study, Al-Kabi et al. created a manually annotated corpus for the standard Arabic (SA). They made it open source with the flexibility in adding and modifying its content. The corpora are divided into five main categories equally, which are: " Technology, Lifestyle (in Food), Sport, Religion, and Economy." It also extended to include the common Arabic dialects: "Arabian Peninsula, Egyptian, Mesopotamian, Maghrebi group, and Levantine" [14]. Al-Twairesh et al. created another corpus by gathering more than 2.2 million tweets. It is manually annotated it, which makes the corpus consist of 17,573 tweets classified into: "positive, negative, mixed, neutral" equally. [3]. Also, "the Qatar Arabic Language Bank Project (QALB)" was a large corpus containing an enormous number of Arabic words in (MSA) and (SA) from six countries, they annotated it manually. They made it open source [15]. In addition to other corpora called MIKA, it was collected and annotated manually by Ibrahim et al., It was classified as positive, negative, and neutral, and is focused on "MSA and Egyptian dialectal Arabic" [16].

As for the focus on vernaculars dialects, there have been many efforts in this area; here, we will investigate the most important of them. The reliance of most researches on collecting data from Twitter, because of its many benefits and flexibility, corresponded to the researchers' need in the data collecting process. One of these researches was by Refaee and Rieser, who collected approximately 8,868 tweets, manually annotated, using several features that positively impact the classification performance [17]. Also, using Twitter data, a corpus called Arap-Tweet has been created that includes many dialects from11 regions and 16 Arab countries. They are manually annotated to classify based on age and gender in addition to the diversity of dialects [18]. Also in another work by Alshutayri and Atwell, they collected data from Twitter along with other sources from Social Media to create a huge corpus of words in vernaculars dialects represented in the five main categories of Arabic dialects: "Gulf, Iraqi, Egyptian, Levantine, and North African" see Fig.1, with more than 266,289 tweets, 9440 comments from news sites, and 812,849 comments from Facebook, which resulted in a corpus containing more than 13,876,504 words [19].

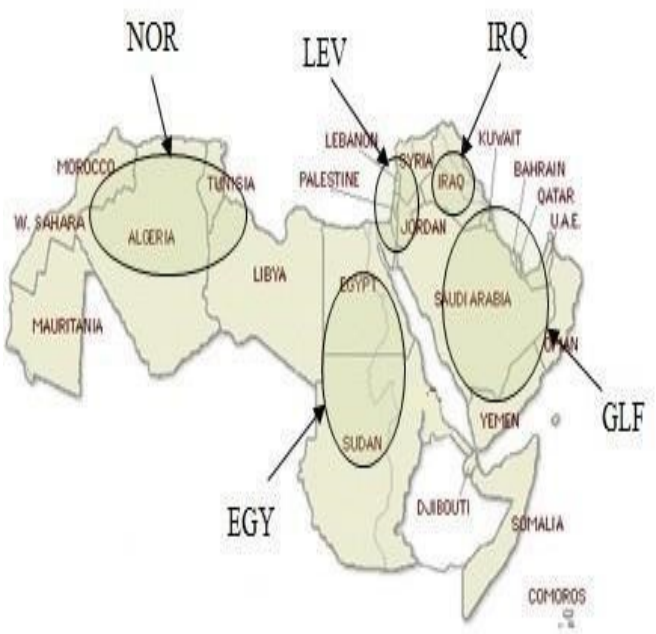

Fig. 1. The five main categories of Arabic dialects.

B. Arabic Topic-based corpora and the techniques and algorithms used

The sentiment analysis algorithm proves its effectiveness in the above researches. Nevertheless, a lot of researchers combine sentiment analysis with topic-based to improve corpus performance and accuracy. The methods of topic-based been introduced with sentiment analysis and prove its effectiveness in various researches. Mike Thelwall and Kevan Buckley in the study [4] apply two new algorithms, lexicon extension, and mood-setting, which results to enhance the performance topic-based lexical sentiment identification for the social media. Two datasets in specific social sciences issues were elected to experiment with these two methods, the UK riot rumors corpus (1,698 tweets), and the AV referendum corpus (17,963 tweets). These two corpora were annotated as a 4-point scale for the sentiment from (strongly negative overall) to (strongly positive overall). In [2], Bing Xiang and Liang Zhou introduced a practical implementation of the Latent Dirichlet Allocation (LDA) algorithm to construct a topic-specific sentiment mixture model. The SemEval2013 corpus has been used in the model experiment, which labels as positive, negative, and neutral. In addition to two data sets, 2 million tweets randomly

selected, and 74 thousand documents from the online newswire. This model shows extra enhancement on the sentiment classification accuracy.

The topic-based sentiment can be very beneficial for research focused around narrow or specific topics. In [20], the researchers Jianfeng Si and others utilize topic-based sentiment from the Twitter sentiment time series to predict the stock market. The Dirichlet Process Mixture (DBM) model has been used to classify topics from the dataset, which were 624782 tweets. And their experiment proves that topic-based models outperform non-topic-based models for stock market prediction. [21] Barkha Bansala and Sangeet Srivastava introduced a new method for Hybrid Topic-Based Sentiment Analysis (HTBSA), to predict elections by utilizing Twitter. The researchers apply the Biterm Topic model (BTM) on more than 300,000 tweets corpus, to identify latent topics, then use pre-existing lexical resources for sentiments polarity. The (HTBSA), results show that tweet can be labeled and weighted with different topics, and this approach can improve elections predications. Furthermore, in a study [22], Ana Reyes-Menendez et al. employs a topic-based sentiment analysis approach to explore the most concern subjects about the environment and public health from Twitter. They collected 5873 tweets using the hashtag (WorldEnvironmentDay), then annotated it by the Support Vector Machine (SVM) algorithm, to negative, positive, and neutral. After that, the tweets been distributed into topics follow to the Sustainable Development Goals (SDGs). The study proves to be helpful to environmental institutions.

However, every year more and more foreign languages corpus is adopting topic-based. [23] Amitava Das and Sivaji Bandyopadhyay build document opinion corpus for the Bengali language from the Bengali newspaper, which contains 2234 sentences. The Theme clustering (k-means) been utilized to identify topic-sentiment aggregation. In [24], Kiet Van Nguyen et al. created a Vietnamese Students' Feedback Corpus (UIT-VSFC) for education and sentiment analysis research. The corpus contains more than 16,000 sentences from students' feedback; its sentiment annotated as natural, negative, and positive. And topic-based annotated for four topics; lecturer, curriculum, facility, and others. The inter-annotator-agreement measures and classification experiments have evaluated the corpus. Moreover, [25] Maria Pontiki et al. continuous creation of SemEval corpus that been created in 2014. This study continues to work on SemEval-2016 based on Aspect Based Sentiment Analysis (ABSA). The corpus contains eight languages and seven different topics.

The majority of Arabic corpus only annotated the sentiment and ignoring topic identification. Nonetheless, there are a few created a topic-aspect sentiment analysis corpus for the Arabic language, such as in [1] researchers Ramy Baly and others build (ARSENTD.LEV), which is multi-topic sentiment analysis corpora for Arabic Levantine. The corpora consist of 4,000 tweets; it annotated as a 5-point scale from very negative to very positive. The Naive Bayes classification has been utilized to gain multi topic-based model. The created model can be used to identify topics and sentiment. [26] A corpora were developed of Arabic hotels' reviews, which is a dataset of the Semantic Evaluation workshop 2016 task 5 (SemEval-ABSA16). The researchers develop the corpora to manage the problems of Aspect Based Sentiment Analysis (ABSA) by apply two methods, support vector machine (SVM) and deep recurrent neural network (RNN). The study clarifies that the SVM method works better than the RNN. And continues in [27] use the same dataset reference (SemEval-2016: Task-5) of Arabic hotels' reviews to improve (ABSA) performance but in this study, the researchers adapt more algorithms; Decision Tree, Decision Tree, Naïve Bayes, K-Nearest Neighbour (KNN), and Support-Vector Machine (SVM).

As been above-mentioned, there are few corpora written in the Arabic language that considers riches annotation, such as topic identification and target of the sentiment. Over and above, it is rare to find a corpus that specific topics concern Saudi Arabia citizens. Therefore, this paper presents Tb-SAC; an Arabic multi topic-based sentiment analysis corpus that contains Saudi tweets, to identify the target of the sentiment and the topic of a text. This corpus can be very beneficial for economic, social, and cultural research for Saudi Arabia citizens.

3. Tb-SAC Corpus

The construction of the (Tb-SAC) corpus, an Arabic multi topic-based sentiment analysis that satisfies the research aim and question. The corpus constructed by following a commonly used approach, such as in [1, 3], we divided the approach to be six main steps. We start by collecting the data from tweets through Twitter API. To perform analysis and prepare the data for annotation and labeling, data cleaning and pre-processing step is achieved, which composed of three phases: cleaning, normalization, and lemmatization. Then we simply perform corpus annotation by the assist of four annotators. Before applying any classification models, we convert the corpus to a readable matrix by text vectorization TF-IDF (term frequency-inverse document frequency) approach. Then eleven classification models been applying on the Tb-SAC corpus. Eventually, the cross-validation method adopted to validate the corpus. Fig. 2. represents the main steps in this corpus development process. The following four sections explain the six steps in more detail.

C. Data collection

We followed a prominent studies approach in creating a corpus, which was adopted on Twitter to collect the dataset. At present, Twitter is an excellent environment for collecting data that helps reach common words among Arabs quickly and accurately. To assemble the data to be used in Tb-SAC Corpus, the main goal was to obtain Saudi tweets only, by the query lang:ar using Twitter API. Therefore, we collected data based on the geographical coordinates of the Kingdom of Saudi Arabia. For more accurate results, we divided Saudi Arabia into five regions, North, South, West, East, and Middle, with radiuses 400km for each region to avoid collecting tweets from the other regions outside Saudi Arabia. The result was with a total of 10,000 tweets collected. Each tweet has the following information: user location, tweet text, and tweet date.

In the collection process, we relied on Python to collect the tweets using tweepy.Cursor, and to determine the geographical areas accurately, we used the geocode parameter on Google Cloab. After reaching 10,000 Arabic tweets from Saudi regions,

we started the initial cleaning process manually intending to exclude the tweets that it collects from outside Saudi Arabia, such as Kuwait and Egypt, so we excluded it manually. Beside eliminate the tweets that have URL links and images since we found most of these tweets is spam. Also excluded tweets containing non-Arabic words or symbols or tweets with unless meaning. After the initial filtering and cleaning process, we reached 4,301 tweets distributed around Saudi Arabia. With 46353 words shown in the Tb-SAC word cloud in Fig. 3.

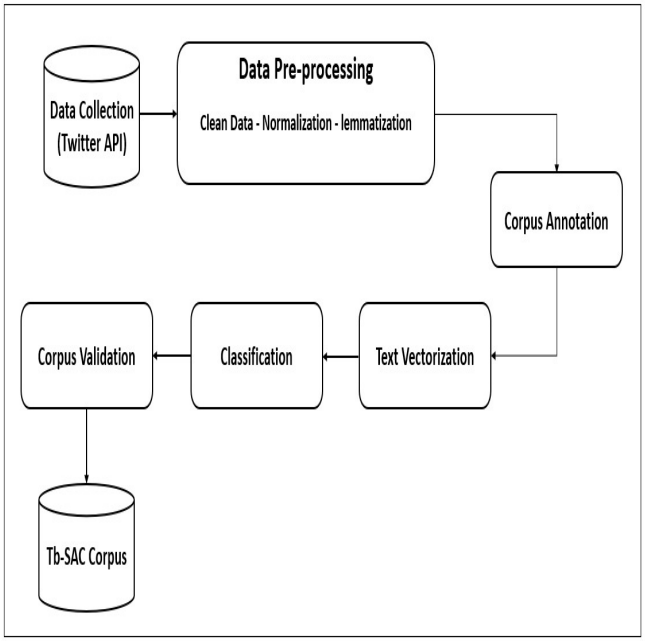

Fig. 2. The framework of Tb-SAC Corpus.

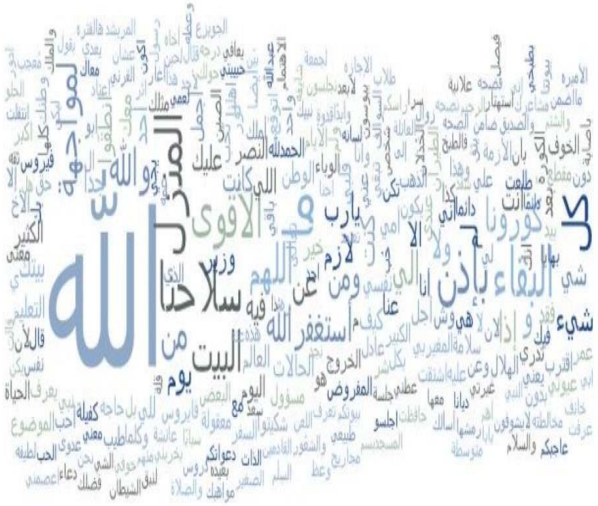

Fig. 3. Word Cloud of Tb-SAC Corpus.

After that, a random sample containing 200 tweets was selected and set as gold set in order to identify general topics for tweets, and then compiling the complete data set based on these topics which are as follows:

Personal tweets were 2178 tweets and were between daily events to express feelings and others (see Fig. 4).

Corona Virus and Quarantine, with 700 tweets (see Fig. 5), the tweets were in large number due to the current situation in which the study was conducted, as the subject prevailing on the majority of social networking sites in the first half of the year 2020 is about Corona Virus.

Religion, with 515 tweets (see Fig. 6).

Entertainment, with 157 tweets (see Fig. 7).

Other, with 751 tweets, which distributed in many topics such as education, economy, sport, food, health, social media, distance working, technology, comedy, and politics. Table 1 will illustrate some examples of tweets indicating each of the five topics.

Table 3 Examples of the topics

Topic	***Arabic Tweet***	***English Translation***
Personal	مابغا شي فحياتي غير اكون مرتاحه ومتطمنه	I don't want anything in my life except to be comfortable and relaxed.
	بس وش سالفة المسابقة	But what is the matter of the competition?
Corona Virus	شابين النار وجالسين بالبيت	We lit the fire and are sitting at home.
	أنا معايا كحة لي ثلاث ايام الله يستر	I have a cough for three days; God bless me.
Religion	الحمدلله ع هالنعمه	Thank God for this blessing.
	الله يكفيك شرهم	May God protect you from their evil.
Entertainment	بلعب معك يلا	I will play with you come on.
	والغريب فالموضوع اللي يلعب معه ينهزم ع طول	The stranger in the matter, with whom he plays, will be defeated immediately.
Other	وشو ذا الجهاز	What is this device?
	مافيه شي المقطع خطأ مبالغين ف الرد واضح تاثير ميولك النصراويه ف الموضوع	There is nothing, the clip is wrong, Exaggerating in the response, the influence of your Nasser team tendencies on the subject is evident

Fig. 4. Word Cloud of Personal Topic.

Fig. 5. Word Cloud of Coronavirus Topic.

Fig. 6. Word Cloud of Religion Topic.

Fig. 7. Word Cloud of Entertainment Topic

STATISTICS OF COLLECTED TWEETS:

Tables 2,3 show the total number of tweets on Tb-SAC Corpus, in addition to the number of tokens for each topic in Table 2 and the sentiment in Table 3.

TWEETS TOPICS:

Table 2. No. of the topics in the TB-SAC CORPUS.

	Personal	***Corona Virus***	***Religion***	***Entertainment***	***Other***	***Total***
Collected tweets	2178	700	515	175	751	4301
No. of Tokens	24763	7148	4894	1800	7930	46535
Size	50.6%	16.3%	12%	3.7%	17.4%	

TWEETS SENTIMENTS:

Table 3. No. of the statements in the TB-SAC CORPUS.

	Positive	***Negative***	***Neutral***	***Total***
Collected tweets	2086	940	1275	4301
No. of Tokens	21271	10521	14743	46535
Size	48.5%	21.9%	29.6%	

D. Data cleaning and pre-processing

Cleaning and pre-processing the dataset is an essential step for the raw tweets because it strongly affects the ability of a model to learn. The pre-processing step contains three phases: remove noise, normalization, and lemmatization. The first phase is a fundamental and essential step to filtrate the dataset

and clarify the tweets for the annotators. In this phase, we remove all the unwanted data in the tweets that make it

ambiguous, such as mentions, hashtags, URLs, white space (extra space and newline), retweets, punctuation marks (!@#$\%^\&*()_+<>?:,;؛ ,}{-؟) and repeated tweets. In the second phase, the normalization phase is the process of standardizing the form of some Arabic letters that have various shapes to be represented in one form without changing the meaning of the word [28]. The dataset normalizes by adopting the tashaphyne library; according to [28], it been proven that it's one of the best normalization libraries for the Arabic language. In this phase, we normalize the Arabic diacritics format letters elongation, and repeated letters. Over and above, we normalize four letters: Ha'a (ه), ya'a (ي), Hamza (ء), and Alef (ا). Table 4 represents some examples of words and the four letters before and after the normalization phase.

Table 4. Example of normalized Arabic text

Normalize	*Before Normalization*	*After Normalization*
the Arabic diacritics	العَربيّة	العربية
letters elongation	بســـــم الله	بسم الله
repeated letters	السلاااااام عليكم	السلام عليكم
Ha'a (ة)	ه, ة	ه
ya'a (ي)	ي, ى	ي
Hamza (ء)	ء, ئ, ؤ	ء
Alef (ا)	أ, ا, إ, آ	ا

In the Lemmatization phase, which use to shrink words to a proper abstract form to be appropriate for the machine learning model [29]. For this phase, we use Farasa API Arabic segmentation, because [30] proves that Farasa is significantly better the state-of-the-art Arabic semesters; Stanford and MADAMIRA, in machine translation and Information Retrieval tasks. And also, they prove that Farasa by orders of magnitude rapid than both using HTTP. Client library in python, we connect to Farasa API and send the dataset for the Lemmatization process. Moreover, using the same Farasa API, we apply the tokenization process on the dataset. Table 5,6, and 7 represents three samples of tweets before and after the three phases of pre-processing.

Table 5. Example of phase 1 data cleaning

Original Tweet	*Phase1 (Clean)*	*Tweet in English translation*
البقاء في المنزل سلاحنا الأقوى لمواجهة بإذن الله #كلنا_مسؤول	البقاء في المنزل سلاحنا الأقوى بإذن الله لمواجهة	Staying at home is our most powerful weapon, God willing, to confront
هل تعرف ما معني أن يثق بك شخص " " خائف دائمًا من الخُذلان	هل تعرف ما معني أن يثق بك شخص خائف دائمًا من الخُذلان	Do you know what it means to be trusted by someone who is always afraid of being let down?
الحمدلله دائما وأبدًا @shmookh_ad	الحمدلله دائما وأبدًا	Thank God ever and forever

Table 6. Example of phase 2 data cleaning

Original Tweet	*Phase2 (Normalization)*	*Tweet in English translation*
البقاء في المنزل سلاحنا الأقوى لمواجهة بإذن الله #كلنا_مسؤول	البقاء في المنزل سلاحنا الاقوى باذن اله لمواجهه	Staying at home is our most powerful weapon, God willing, to confront
هل تعرف ما معني أن يثق بك شخص " " خائف دائمًا من الخُذلان	ان هل تعرف ما معنييثق بك شخص خاءف داءما من الخذلان	Do you know what it means to be trusted by someone who is always afraid of being let down?
الحمدلله دائما وأبدًا @shmookh_ad	الحمدله داءما وابدا	Thank God ever and forever

Table 7. Example of phase 3 data cleaning

Original Tweet	*Phase3 (Lemmatization)*	*Tweet in English translation*
البقاء في المنزل سلاحنا الأقوى لمواجهة بإذن الله #كلنا_مسؤول	بقاء', 'في', 'منزل', 'سلاح', 'اقوي', 'اذن', 'اله', 'مواجه	Staying at home is our most powerful weapon, God willing, to confront
هل تعرف ما معني أن يثق بك شخص " " خائف دائمًا من الخُذلان	هل', 'عرف', 'ما', 'معني', 'ان', 'وثق', 'ب', 'شخص', 'خاءف', 'داءم', 'من', 'خذلان	Do you know what it means to be trusted by someone who is always afraid of being let down?
الحمدلله دائما وأبدًا @shmookh_ad	حمدله', 'داءم', 'ابدا'	Thank God ever and forever

E. Annotation Methodology

In the annotation process for this corpus, the annotation carried out using In-house labeling, which was implemented by a group of four volunteers who were postgraduate students in various study fields. The annotators used the annotation process to develop a set of a tagged corpus to analyze topics for all 4301 tweets. The internal annotators asked to assign topics to each Tweet, and they also asked to identify the sentiments included in each Tweet using a three-point scale with the following designations: positive, negative, neutral.

- Positive: For all tweets that contain positive feelings, whether expressed explicitly, such as words indicating positive words, such as "سعيد", "خبر مفرح", or expressed implicitly.
- Negative: Tweets that contain negative words like " سيء", "جدا", "ما اعجبني", or that include some words hurt to another party are classified as negative tweets whether the expression is explicit or implicit.

- Neutral: It used to express tweets that do not contain certain feelings expressed as tweets that contain only news or Quranic verses. Sentiments are illustrated in Table 8.

The annotators started to classify all tweets based on their first impressions of each tweet they read. Finally, Since the emotions that included in some tweets have a great impact in identifying the feelings for the tweet, and because we have removed the tweets that contain them, we have taken care during the manual cleaning process to choose the tweets that are not affected by the removal of the emojis, emotions, etc. in order to make sure to obtain accurate results for the annotation process.

Table 7. Example of statements in Tb-SAC.

Labels	*Arabic Tweet*	*English Translation*
Positive: if there are any positive words, such as in this tweet.	فكرة البرنامج بطله وجميله السؤال لما يجي منطفل مره يكون شي لطيف	The idea of the program is heroic and beautiful. When the question comes from a child, it is a very nice thing
Negative: if there are any negative words, such as in this tweet.	احس الرياض حزينة جد أ	I feel that Riyadh is very sad
Neutral: if there is no clear sentiment in the tweet.	وزير التعليم البقاء في المنزل سلاحنا الأقوى بإذن الله لمواجهةفايروس كورونا	The Minister of Education, staying at home, is our most powerful weapon, God willing, to confront the Coronavirus

According to the annotation process for sentiments, the mode for the classified sentiments measured for the four annotators, and it concluded that the majority of sentiments were positive at 47.15 %, neutrality at 33.15 %, and the lowest was the negative tweets with 19.70 %. Table 9. Shows the number and percentage of each annotator in the sentiment's annotation process.

Table 8. Annotation Process of sentiments for Each Annotator.

	Annotator 1	*Annotator 2*	*Annotator 3*	*Annotator 4*	
Positive	2759	1500	2086	1817	47.15%
Negative	1076	554	940	710	19.70%
Neutral	466	2247	1275	1774	33.15%

On the other hand, after the analysis of the gold set, we identified the topics mentioned previously, which were personal, coronavirus, religion, entertainment, and others. Annotators also asked to classify the tweets based on these topics. As shown in Table 10. Also, Table 11. provides examples from the Tb-SAC corpus for each topic.

Table 9. No. of topics Tb-SAC.

Topic	*Size*	*Sub-topics*
Personal	1936	Personal opinions, daily events
Coronavirus	778	Coronavirus news, Quarantine
Religion	541	Quranic verses, hadith, and religious advice
Entertainment	186	Electronic games, movies, series, and TV shows
Other	860	education, health, food, sports, social media, distance working, comedy, political, technical and economy

Table 10. Example of Topics in Tb-SAC.

Topic	*Arabic Tweet*	*English Translation*
Personal	اتوقع على حسب البيئه اللي حوله وممكن هو يصنع ثقافته من الكتب	I expect according to the environment around him and its possible that he makes his culture from books
Coronavirus	ارتفاع عدد الوفيات بفيروس 21كورونا في تركيا إلى	Corona virus death toll in Turkey increased to 21
Religion	اللهم ص ل وسلم على نبينا محمد	God blessing and peace upon our Prophet Muhammad
Entertainment	اعطوني العاب بالجوال غير ببجي	Give me mobile games other than Pubg
Other	لاعب ممتاز مكانه في الهلال عشان يحقق البطولات	An excellent player who takes his place in Al Hilal to achieve the championships

To ensure the accuracy of the annotation process and to ensure the reliability of the selected annotators, kappa coefficients were calculated for the agreement degree between the four annotators. As the results indicate, the Fleiss' kappa agreement for the four annotators was 0.42, which is considered moderate based on [31].

F. Experiments and validation

In this section, we introduce the results of applying multiple classification models on the Tb-SAC corpus. We also conducted several cross-topic experiments to emphasize the effects of different percentages between the topics in the corpus.

The Tb-SAC corpus was split into 80% for the training set and 20% for the test set. And we convert the corpus into a matrix to be readable by the classification models. The matrix is constructed by the term frequency-inverse document frequency (TF-IDF) method with N-gram, by adopting the Tfidf-Vectorizer library in python. We generated three matrixes for uni-gram, bi-gram, and tri-gram. Before we apply any classifier, the Linear SVC model applies to the three matrixes to get the best fit hyperplane. Then, we apply eleven classification models on the training set of the corpus; Table 12 represent the eleven classification models, and the results of the accuracy score in each N-gram. The results show that RBF Support Vector Machines has the highest accuracy score among other models, with a score of 60.86%.

Table 11. The Eleven Classification Models Results.

Classifier	*Uni-gram*	*Bi-gram*	*Tri-gram*
Linear Support Vector Machines	**59.81**	**59.70**	**59.70**
RBF Support Vector Machines	60.86	60.86	60.86
k-Nearest-Neighbors	43.67	44.25	44.25
Naive Bayes	58.89	58.89	58.89

Logistic Regression	60.39	60.63	60.63
Gradient Boosting	57.72	57.14	57.49
Random Forest	56.91	56.56	58.19
AdaBoost	56.79	55.17	55.17
Decision Tree	51.22	52.96	50.87
Bernoulli Naive Bayes	59.12	58.77	58.77
Stochastic Gradient	59.70	58.77	58.65

For the validation step, we perform a cross-validation method. And we only consider the classification models that get results more than 59%, which are Linear Support Vector Machines, RBF Support Vector Machines, Logistic Regression, Bernoulli Naive Bayes, and Stochastic Gradient. Thus, we apply a 5-fold cross-validation and 10-fold cross-validation. Table 13,14 show the 5-fold cross-validation and 10-fold cross-validation results for each N-gram, where RBF Support Vector Machines and Logistic Regression have the highest result, both around 58%. However, there is a slightly higher accuracy in 10-fold cross-validation results.

Table 11. 5-Fold Cross-Validation Results.

Classifier	*Uni-gram*	*Bi-gram*	*Tri-gram*
Linear Support Vector Machines	57.54	57.43	57.43
RBF Support Vector Machines	58.08	58.27	58.27
Logistic Regression	**58.36**	**58.47**	**58.47**
Bernoulli Naive Bayes	57.13	57.03	56.92
Stochastic Gradient	57.73	57.66	57.52

Table 12. 10-Fold Cross-Validation Results.

Classifier	*Uni-gram*	*Bi-gram*	*Tri-gram*
Linear Support Vector Machines	57.59	57.68	57.66
RBF Support Vector Machines	**58.52**	**58.59**	**58.59**
Logistic Regression	58.27	58.24	58.24
Bernoulli Naive Bayes	57.10	57.13	56.87
Stochastic Gradient	58.24	57.99	57.99

Finally, we perform several cross-topic experiments to show the effects of having considerable differences in percentages of the topics in the corpus. We apply 10-fold cross-validation with RBF Support Vector Machines and Logistic Regression for cross-topic experiments as it proves above that it generated higher results. For the experiments, we perform three tests, each with two topics: (coronavirus, personal), (Religion, personal), and (entertainment, personal). We combine the personal topic in each test because it has the highest percentage in the corpus. Table 15 shows the three test results. Test that has the most differences between topics percentage, which are entertainment and personal topics get the highest accuracy among other tests with an accuracy of around 90%, that consider as overfitting. That proves the models perform better if they are no huge differences between topics percentage in the corpus.

Table 13> Cross-Topics Tests Results.

Classifier	*Coronavirus & personal*	*Religion & personal*	*Entertainment & personal*
RBF Support Vector Machines	**87.44**	**83.79**	90.25
Logistic Regression	87.41	83.04	**90.31**

4. Research Challenges

In this section, we introduce obstacles and challenges faced by the research regarding Arabic language, availability of data, problems in annotation, or in the process of developing the Tb-SAC corpus overall. In the data collection phase, we collected the tweets using geocode parameters to collect tweets according to a certain zone, as mention before. The problem is that the five-zone that we used was not very specific for Saudi Arabia, where the problem mostly in west and east zones. Thus, we get several tweets that are from neighboring countries such as Kuwait, Emirates, Iraq, Egypt. Moreover, a huge number of tweets contents about coronavirus and quarantine since the time we collected the data was at the beginning of the spread coronavirus in Saudi Arabia. So, it has a significant impact on the chosen topics and the corpus overall.

The process of pre-processing Arabic language data is complicated, as the nature of the Arabic language is complex, and it lacks text mining tools. Also, finding libraries in python for normalization and lemmatization was very challenging. Due to the Arabic language nature, not all Arabic libraries generate valid results, especially for Arabic dialects, where some Arabic libraries change the whole meaning of the word in the lemmatization phase.

In the annotation phase, we choose four annotations, and most of the annotation work done manually, as mention above. The manual process for the four annotators took the most time of the corpus developing process. Also, in sentiment labeling, some annotators were very subjectivity.

And Annotators express that they faced some challenges in labeling topics for tweets since some tweets may refer to two topics. For example, بنجلس نلعب ببجي الى مايخلص الكارونا, Translation: we will play PUBG until the coronavirus disappear. Some annotations label the tweet with coronavirus topic and other with entertainment topic.

5. Conclusion

In this paper, we presented the Tb-SAC, a corpus for topic-based using Twitter in Saudi dialect tweets. Based on the manual conducted annotation on the gold set of 200 tweets, which retrieved using Twitter API from Saudi regions. The annotation process performed by postgraduate students, and it includes determining a suitable topic for each tweet from the topics that were previously identified (personal, coronavirus, religion, entertainment, other), besides deciding the tweet sentiment (positive, negative, neutral). Consequently, the presented corpus consists of 4301 tweets collected from Saudi dialects. Finally, we applied eleven different classification models on the corpus to compare the result of the classifiers in order to get excellent accuracy rates. Also, the validation process was performed by the cross-validation model, where RBF Support Vector Machines and Logistic Regression have the highest result, which was around 58% in 10-fold cross-validation.

In the near future, we hope to present the Tb-SAC corpus as an open work of the scientific research community. We will also make it subject to amendment and development by researchers in this field.

References

[1] Baly, R.; Khaddaj, A.; Hajj, H.; El-Hajj, W.; Shaban, K.B. ArSentD-LEV: A multi-topic corpus for target-based sentiment analysis in Arabic levantine tweets. arXiv preprint arXiv:1906.01830 2019.

[2] Xiang, B.; Zhou, L. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 434–439.

[3] Al-Twairesh, N.; Al-Khalifa, H.; Al-Salman, A.; Al-Ohali, Y. Arasenti-tweet: A corpus for Arabic sentiment analysis of Saudi tweets. Procedia Computer Science 2017, 117, 63–72.

[4] Thelwall, M.; Buckley, K. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology 2013, 64, 1608–1617.

[5] Al-Twairesh, N.; Al-Khalifa, H.; Al-Salman, A. Towards analyzing Saudi tweets. 2015 First International Conference on Arabic Computational Linguistics (ACLing). IEEE, 2015, pp. 114–117.

[6] Nasukawa, T.; Yi, J. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd international conference on Knowledge capture, 2003, pp. 70–77.

[7] Liu, B. Sentiment analysis: Mining opinions, sentiments, and emotions; Cambridge University Press, 2015.

[8] Sawalha, M.; Alshargi, F.; Alshdaifat, A.; Yagi, S.; Qudah, M.A. Construction, and Annotation of the Jordan Comprehensive Contemporary Arabic Corpus (JCCA). Proceedings of the Fourth Arabic Natural Language Processing Workshop, 2019, pp. 148–157.

[9] Zaghouani, W. Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835 2017.

[10] Al-Sulaiti, L.; Atwell, E.S. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics 2006, 11, 135–171.

[11] Saad, M.K.; Ashour, W. OSAC: Open source Arabic corpus. Proceedings of the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, 2010, pp. 557–562.

[12] Abbas, M.; Smaili, K. Comparison of topic identification methods for Arabic language. Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP, 2005, pp.14–17.

[13] Abdul-Mageed, M.; Diab, M.T. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. LREC, 2012, Vol. 515, pp. 3907–3914.

[14] Al-Kabi, M.; Al-Ayyoub, M.; Alsmadi, I.; Wahsheh, H. A prototype for a standard arabic sentiment analysis corpus. Int. Arab J. Inf. Technol. 2016, 13, 163–170.

[15] Mohit, B.; Rozovskaya, A.; Habash, N.; Zaghouani, W.; Obeid, O. The first QALB shared task on automatic text correction for Arabic. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 39–47.

[16] Ibrahim, H.S.; Abdou, S.M.; Gheith, M. MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis. 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS). IEEE, 2015, pp. 353–358.

[17] Refaee, E.; Rieser, V. An arabic twitter corpus for subjectivity and sentiment analysis. LREC, 2014, pp.2268–2273.

[18] Zaghouani, W.; Charfi, A. Arap-tweet: A large multi-dialect twitter corpus for gender, age, and language variety identification. arXiv preprint arXiv:1808.07674 2018.

[19] Alshutayri, A.; Atwell, E. A social media corpus of Arabic dialect text. Computer-Mediated Communication and Social Media Corpora. Clermont-Ferrand: Presses Universitaires Blaise Pascal 2019.

[20] Si, J.; Mukherjee, A.; Liu, B.; Li, Q.; Li, H.; Deng, X. Exploiting topic based twitter sentiment for stock prediction. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013, pp. 24–29.

[21] Bansal, B.; Srivastava, S. On predicting elections with hybrid topic based sentiment analysis of tweets. Procedia Computer Science, 2018, 135, 346–353.

[22] Reyes-Menendez, A.; Saura, J.R.; Alvarez-Alonso, C. Understanding#WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach. International journal of environmental research and public health 2018, 15, 2537.

[23] Das, A.; Bandyopadhyay, S. Topic-based Bengali opinion summarization. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010, pp. 232–240.

[24] Van Nguyen, K.; Nguyen, V.D.; Nguyen, P.X.; Truong, T.T.; Nguyen, N.L.T. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. 2018 10th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2018, pp. 19–24.

[25] Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; Al-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; others. Semeval-2016 task 5: Aspect based sentiment analysis. 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016.

[26] Al-Smadi, M.; Qawasmeh, O.; Al-Ayyoub, M.; Jararweh, Y.; Gupta, B. Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews. Journal of computational science, 2018, 27, 386–393.

[27] Al-Smadi, M.; Al-Ayyoub, M.; Jararweh, Y.; Qawasmeh, O. Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. Information Processing & Management 2019, 56, 308–319.

[28] Alotaibi, S.S.; Anderson, C.W. Extending the knowledge of the Arabic sentiment classification using a foreign external lexical source. Int. J. Nat. Lang. Comput, 2016, 5, 1–11.

[29] El-Shishtawy, T.; El-Ghannam, F. An accurate arabic root-based lemmatizer for information retrieval purposes. arXiv preprint arXiv:1203.3584 2012.

[30] Abdelali, A.; Darwish, K.; Durrani, N.; Mubarak, H. Farasa: A fast and furious segmenter for arabic. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016, pp. 11–16.

[31] Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. biometrics, 1977, pp. 159–174.