SDCT: Multi-Dialects Corpus Classification for Saudi Tweets

Afnan Bayazed[1], Ola Torabah[2], Redha AlSulami[3], Dimah Alahmadi[4], Amal Babour[5], Kawther Saeedi[6]
Information Systems Department, King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—There is an increasing demand for analyzing the contents of social media. However, the process of sentiment analysis in Arabic language especially Arabic dialects can be very complex and challenging. This paper presents details of collecting and constructing a classified corpus of 4180 multi-dialectal Saudi tweets (SDCT). The tweets were annotated manually by five native speakers in two stages. The first stage annotated the tweets as Hijazi, Najdi, and Eastern based on some Saudi regions. The second stage annotated the sentiment as positive, negative, and natural. The annotation process was evaluated using Kappa Score. The validation process used cross validation technique through eight baseline experiments for training different classifier models. The results present that the 10-folds validation provides greater accuracy than 5-folds across the eight experiments and the classification of the Eastern dialects achieved the best accuracy compared to the other dialects with an accuracy of 91.48%.**

Keywords—Arabic dialects; dialects classification; language classification; natural language processing; Saudi dialects; sentiment analysis; Twitter

I. Introduction

Today, there are roughly 6500 spoken languages around the world, and each language involves different multiple dialects [1]. Arabic language is one of the most used languages in the world. Arabic is the official language of 22 countries, and it is spoken by over 400 million people. It is considered the fourth language used the most on the Internet [2]. There are three varieties of Arabic language which are Classical Arabic (CA), Modern Standard Arabic (MSA) and Arabic dialects (AD). The CA is a form of Arabic language used in literary texts and the Quran (Islam's Holy Book). The MSA is the essential Arabic form that is used commonly in formal conversations, media, education, newspapers, magazines, and formal TV programs. The AD is used in informal communication, and it is divided by geographical region [3]. The AD geographical regions are Egyptian, North Africa, Levantine, Iraqi, and Gulf [1]. However, the Gulf region consists of six countries: Saudi Arabia, United Arab Emirates, Qatar, Kuwait, Bahrain, and Oman, where each country has its own dialect. As for Saudi Arabia, also each different region has its own dialect. In Saudi Arabia, the dialects are Hijazi in the western region, Najdi in the Middle region, Southern dialect in the Southern region, Northern dialect in the Northern region, and Eastern dialect in eastern region. The AD has huge differences between them that can be considered different languages; therefore, Arabic language and its dialects required further intensive study and analysis. Most of Arabic Natural Language processing (NLP) applications are dedicated to the MSA like sentiment analysis, machine translation, speech recognition, and speech synthesis. Moreover, the Arabic NLP tools such as part-of-speech (POS) tagging, morphological analysis, and disambiguation are designed specifically for MSA, and for that, it gave a less accurate result for AD.

The Arabic NLP resources are focused on the MSA that has covered all orthographic varieties and have a rich morphology, and a strong syntactic system. As for the AD, the Arabic NLP resources do not cover it as well as the MSA. Furthermore, the AD is spoken languages with no writing system. Creating resources for the Arabic dialects is challenging in the Arabic NLP but it is necessary [4-6].

Particularly with the proliferation amounts of textual data on social media websites and microblogs, such as Twitter and Facebook, there is a huge resource for the Arabic dialects. Social media is an important communication tool for people to write about their daily life, share information, add reviews or opinions, explore the latest news and search for real-time news events. Arabic users tend to communicate with each other using the unstructured and ungrammatical slang Arabic language. Twitter is one of the world's most popular platforms for internet users. Twitter users send about 500 million tweets per day, where each tweet contains 280 characters [7]. The Arab people have been influenced by the recent evolution in technology. The total number of Arabic users on Twitter are more than 11 million, with 27.4 million tweets per day. The most active users are from Saudi Arabia with about 30% of all the tweets [8].

Al-Twairesh et al. in [9] claims that the lack of Arab corpora is one of the challenges facing a sentiment analysis of Arab. Accordingly, this research aims to utilize the huge Arabic textual data and prepare it as language resources for the Saudi dialects. This paper's contributions can be summarized as follows:

- Build Saudi Dialects Corpus from Twitter called SDCT and make it available as an open source for the research community.
- Classify SDCT depending on different Saudi dialects (Hijazi, Najdi and Eastern).
- Provide sentiment labelling of each dialect mostly to Positive, Negative and Neutral.

The paper is organized as follows: The previous related work is described in Section II. Section III explains the methodology for creating the corpus and its annotation, includes a preview of the twitter corpus collection and Saudi tweet analyzes, and discusses the experimental findings. Section IV presents the challenges of this research. Finally, the conclusion and future recommendations are shown in section V.

II. Related Works

For providing a comprehensive overview, we survey the related works available in Arabic corpora and dataset. Several studies have proposed number of approaches in the Arabic dialects classification. Also, there are enormous studies conducted in sentiment analysis for Arabic dialects.

For the purpose of creating frameworks for sentiment analysis, Duwairi et al. in [10] developed a framework for sentiment analysis on Arabic Tweets in text reviews. They used a translated version of English lexicon called SentiStrength and extended it with synonyms list for every word in the lexicon as a seed list. The polarity for each word in the seed list is expressed as -1 for negative sentiment and 1 for positive sentiment. They used a set of 4400 Arabic tweets, where each tweet was tokenized into terms and the sentiment of the tweet was determined by summing the scores of all the terms in the tweets where the sentiment of the tweet was considered positive if its summation is greater than 0, negative if its summation is less than zero, and neural if the summation equals zero. For the performance of the proposed framework, they applied two experiment without and with stemming the tweets. The results showed that the framework achieved good results and improved the precision, recall, accuracy and reduced error rate with stemmed tweets. Duwairi et al. in [11] proposed a framework for Arabic text sentiment analysis based on a created crowdsourcing API to manually annotate a training dataset of 25000+ tweets as positive, negative, or neutral. To test the performance of the framework, they used Rapidminer built-in classifiers named Naïve Bayes (NB), k-nearest classifier (KNN), and Support Vector Machines (SVM) on a stratified sample of 1000 tweets from the training dataset. For each classifier, the applied two experiments without and with stopwords/ stemming. The result showed that the best accuracy was achieved by SVM when no stopwords and no stemming were used.

For creating corpus of Arabic sentiment analysis, Refaee and Rieser in [12] constructed a corpus supporting Subjectivity and sentiment analysis (SSA) and collected 8,868 Arabic twitter feeds from multiple Arab dialects. Then, manual annotation processes have been performed by two annotators to polar, positive, negative, neutral and mixed. Furthermore, Nabil et al. in [13] presented an Arabic social sentiment analysis dataset (ASTD) consisting of 10K tweets that were manually annotated using Amazon Mechanical Turk (AMT) to objective, subjective positive, subjective negative, and subjective mixed. Assiri et al. in [14] created the first Saudi annotated corpus collected randomly from Twitter trending hashtags promotion in Saudi Arabia. In accordance with the preprocessing and manual annotation, they collected around 4700 tweets. The dataset has manually annotated based on the sentiment text polarity as positive, negative, and neutral using an application user interface. Furthermore, there is a similar effort in [15] where Altwairesh et al. presented and produced a comprehensive corpus of 18K tweets by using a specific annotation system.

For creating corpus of Arabic dialects and language, Alshutayri and Atwell in [16] built a corpus of 13.8M collected from Twitter, newspapers, and Facebook. The corpus was annotated into five different dialects: Egyptian, Gulf, Levantine, Iraqi, and North African. For the annotation process, they developed an online game via a website where players can involve in the annotation by classifying their dialects. Likewise, Mubarak and Darwish in [17] created a large corpus of Arabic dialects collected from the Twitter platform. The size of the corpus is 175M. The corpus was annotated into six different dialects: Saudi Arabian, Egyptian, Algerian, Iraqi, Lebanese and Syrian. For the annotation process, native speakers of each dialect have involved in determining if a tweet belongs to their dialects or not. Alshutayri and Atwell in [1] created a corpus of 210,915K tweets containing five Arabic dialects: Gulf, Iraqi, Egyptian, Levantine, and North African. For the annotation process, they used Waikato Environment for Knowledge Analysis (WEKA). Altamimi et al. in [18] created the corpus containing 122K tweets for Arabic dialects collected from twitter. Tweets were annotated manually into five labels: Gulf; Egyptian; Levantine; Maghrebi; and Iraqi; in addition to Modern Standard Arabic (MSA) and Classical Arabic (CA). Maghfour et al. in [19] centered their study on classified the Facebook comments as expressed in MSA or in Moroccan Dialect besides the Sentiment analysis (SA) classification in comments. Hence, they performed two different schemas. The first one is a classical schema that considered all Arabic dialects and languages as homogeneous thus, then they applied sentiment analysis on the collected dataset at once. In the second schemas, they proposed to classify the Arabic language into two sub-dataset MSA and Dialect Arabic (DA) beforehand sentiment analysis. Therefore, they applied different preprocessing and Arabic dialect stemmers on each sub-dataset. In supervised classification, they employed the most two reported sentiment classification algorithms, Naive Bayes (NB) and Support Vector Machine (SVM). In the testing phase, the four combinations of weighting schemes n-gram and extraction schemes have been utilized. This study has recorded a high score in the classical schema with the NB algorithm. The Similar effort presented by Medhaffar et al. [20] where they developed the Tunisian dialect dataset that composited 17K comments collected from Facebook. They applied three classification algorithms SVM, NB, and multi-layer perceptron MLP. Their models have shown better accuracy than other models that trained on MSA.

Accordingly, there are enormous studies that provide a public Arabic dialects' lexicon and corpus to address complexity and difference in Arabic language and its dialects. Furthermore, some of the researches have specialized to study sentiment analysis in Saudi dialects text aligns with the increasing demand to analyze social media content in the Saudi market. However, the contributions related to Saudi dialects are still insufficient and limited. In addition, there is no corpus that classified the different Saudi dialects according to the regions. Besides no sufficient research that study sentiment analysis in each specific classified Saudi dialects. Therefore, we seek to center our work in creating a public corpus of sentiment analysis and classification of Saudi dialects.

III. Experiment

This section illustrates our approach for building a SDCT corpus that is dedicated to Saudi dialects. First, we collected the data and targeted Twitter as the main source of data collection. Then we conducted the preprocessing phase that involved three main tasks, which are data cleaning, normalization, and lemmatization. After the data collection and preprocessing, the data were manually annotated, hence the annotation was evaluated using Kappa Score [24]. In addition, we extracted the features to be used in the training set. Then the classification was conducted by the classifiers and, finally, we validated the classifiers via the cross-validation technique. Fig. 1 illustrates the proposed approach for building the SDCT corpus.

A. Data Collection

This research aims to build a SDCT Corpus for Saudi dialects collected from the social network application Twitter. Initially, we planned to cover all five of the dialects in Saudi Arabia: Hijazi, Najdi, Southern, Northern, and Eastern. However, according to our study of tweets in Saudi Arabia we observe that the southern and northern dialects are not widely used, and the number of tweets in these dialects are scarce compared to the Najdi, Hijazi and Eastern dialects. Hence, we limited our experiment and focused only on the three main dialects: Hijazi, Najdi, and Eastern, which are mostly used on Twitter in Saudi Arabia. In the data collection process, we used Twitter API for developers, which is provided by Twitter to allow access to their social media content, and ‘Tweepy’, a Python library for retrieving tweets. Retrieved tweets were stored in the CSV format and, therefore, could be accessed using an Excel spreadsheet. Furthermore, we retrieved approximately 8923 tweets and reached a total of 4181 after the cleaning process, as illustrated in Table I. In addition, we mainly relied on the terms used particularly in each specific dialect and used both the time zone of the dialect region and trending hashtags, as illustrated in Table II. The collection process was accomplished in around two weeks, from March 15th, 2020 to March 28th, 2020. In addition, the corpus is available upon any request from the authors for research and testing.

B. Data Preprocessing

The preprocessing phase is one of the important steps in text mining. It prepares the raw text for the next phase by removing unwanted or annoying data and reduces the dimensionality size of text data as well as normalizing the text.

Therefore, in SDCT corpus, we divided the data preprocessing into three phases, which are the data cleaning phase, normalization phase, and finally, the lemmatization phase. Firstly, we manually removed Ads tweets in the data cleaning phase, the tweets that are not related to any of the dialects that we identified such as Lavanteen or EGY dialects, and unhelpful short tweets, such as تمامwhich means "OK", and كيف الحالwhich means "How are you". Then we automatically removed by coding noise data, such as links (http://, https://), emoji, mentions (@Username), retweets, hashtags as (#corona , # كورونا), and punctuation (!@#$%&ˆ *()_ +<>?:,;-{}c,c) from our SDCT corpus. Secondly, in the normalization phase, we used the Tashaphyne library for the normalization process [21]. This included normalizing letters, such as Alef (أ ، ا ، إ ، آ), Hamza (ء ، ؤ ، ئ), Ya’a (ي ، ى) and Ha’a (ة ، ه), strip repeated letters, elongation (Tatweel), and diacritics (Tashkeel and Harakat). The normalization process is illustrated in Table III.

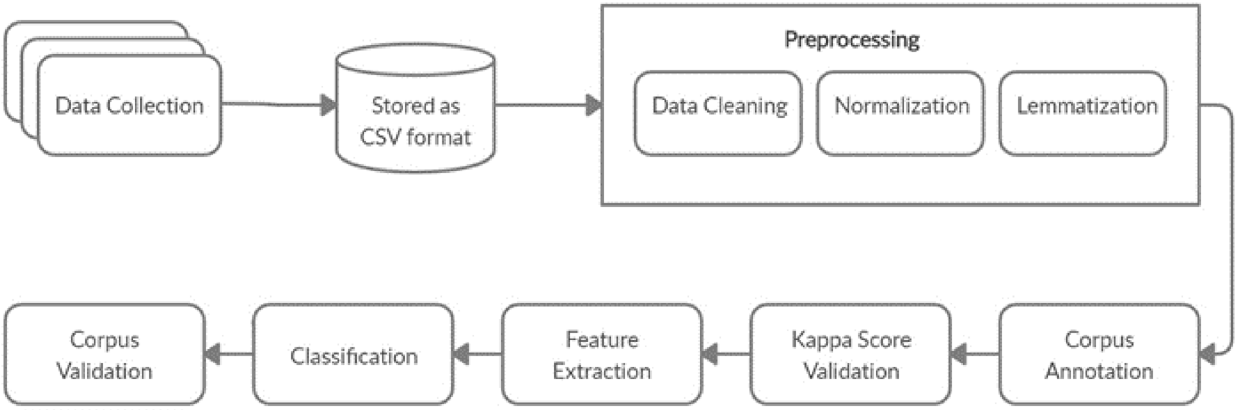

Fig. 1. The Proposed Approach for Building a SDCT Corpus.

TABLE I. This Size of Collected Tweets before and after the Cleaning Process

Prosperities		Data Collection		After Cleaning	
Dialect	***Collection Process***	***Size***	***Total***	***Size***	***Total***
Hijazi	Time Zone, Keyword	3543	8923	1507	4181
Najdi	Trending Hashtag, Keyword	3450		1341	
Eastern	Keyword	1930		1333	

TABLE II. EXAMPLES OF TWEETS RETRIEVED BY USING THE TIME ZONE, KEYWORDS AND TRENDING HASHTAGS

Dialects	Keyword, Time zone, Hashtag	Example (Arabic)	Example (English Translation)
Hijazi	دحين "Daheen" أهرج "Ahrej" مرة "Marra" كمان "Caman"	دحين كيف مصدقة نفسي اني بقوم الساعة 8 ااهرج ماعليك نسمعك اول مرة اشوف الكلمة أصلاً اشتقت للدوام انا كمان	Now, how do I believe myself that I'm going to wake up at 8 Speak, do not worry we hear you Basically, it's the first time I see the word I missed work too
Najdi	عيا" Aiyah" هالحين "Halheen" وش "Wesh" امحق " Emhag"	عطني رابطه تكفا حسابي عيا يفتح يقول حمل الجديد والله كنا عايشين احسن من هالحين والدنيا سهالات وش دا الهجوم لا تدعي على احد كريمة بحق غيري امحق كرم	Give me its link please, my account can't open, it asks to download the new version We were living better than now, and the world was easy What is this attack, do not claim anyone Generous to others, great generosity
Eastern	وايد "waeed" اشفيج"eshfej" شنو "shno" صج "sij"	احسك تحب كرة القدم وايد شفيج معصبه عيوني ترا نمزح شنو اسم اللعبة هذا اللي اشوفه صج ولا جرافيكس	I feel you love football very much Why are you angry we are joking What is the name of the game? This is what I see is true or graphics?

TABLE III. NORMALIZATION PROCESS

Example	Replaced By	Process Name
صدمممممتتتننيييي	صدمتني	Strip repeated letters
ســـــــــــلام	سلام	Strip elongation (Tatweel)
ضيّقت	ضيقت	Strip diacritics (Tashkeel and Harakat)
أ إ آ ا	ا	Normalizing Alef
ي ى	ي	Normalizing Ya'a
ة ه	ه	Normalizing Ha'a
ئ ؤ ء	ء	Normalizing Hamza

Lastly, we preferred to use Farasa tool, one of the NLP and morphological tools for the lemmatization process. This is a fast and accurate new Arabic segmenter that uses SVM for ranking and is proposed by [22]. Table IV illustrates some examples of lemmatization by using Farasa. Furthermore, Farasa had been used by some previous researches, where it made fewer stemming errors. Some examples of Preprocessing phases are shown in Fig. 2.

TABLE IV. EXAMPLE OF LEMMATIZATION PROCESS USING FARASA

Word	Lemma	English Translation
كتبنا	كتب	Wrote
يشتكون	اشتكى	Complained
قلنا	قال	Said
توقعنا	توقع	Expected
مكتبئة	مكتئب	Depressed

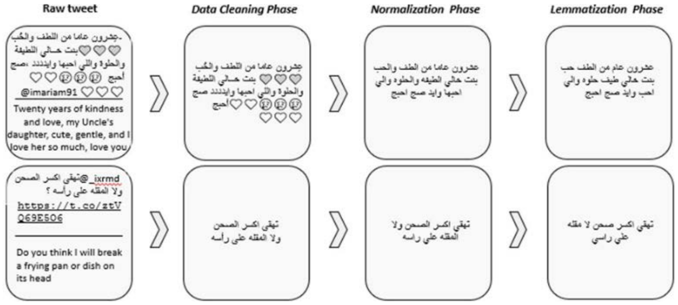
Raw tweet
Data Cleaning Phase
Normalization Phase
Lemmatization Phase
عشرون عاما من اللطف والحُب ♡♡♡بنت خــالي اللطيفة والحلوة واللي احبها وايددد ،صح احبج
@imariam91 ♡♡♡
Twenty years of kindness and love, my Uncle's daughter, cute, gentle, and I love her so much, love you
عشرون عاما من اللطف والحُب ♡♡♡ بنت خــالي اللطيفة والحلوة واللي احبها وايددد صح احبج♡♡
♡♡♡
عشرون عاما من الطف والحب بنت خالي الطيفه والحلوه والي احبها وايد صح احبج
عشرون عام من الطف حب بنت خالي طيف حلوه والي احب وايد صح احبج
@_ixrmd تبغى اكسر الصحن ولا المقله على رأسه ؟
https://t.co/ztVQ69E5O6
Do you think I will break a frying pan or dish on its head
تبغى اكسر الصحن ولا المقله على رأسه
تبغي اكسر الصحن ولا المقله علي راسه
تبغي اكسر صحن لا مقله علي راسي

Fig. 2. Examples of Preprocessing Phases.

C. Annotation Methodology

Annotation is the process of adding text for labels that rely on both the classification output and sentiment [15, 23]. In this study, we inducted five annotators who graduated from King Abdulaziz University and are considered to be native speakers of Saudi Arabian. Hence, in the first stage, four of the annotators manually annotated 4,181 tweets corpus for both dialects and sentiment. Firstly, every annotator classified each tweet dialect into one of the following four labels: Hijazi, Najdi, Eastern, and Saudi dialect (SA) for the unknown or the mixed dialect tweets. After that, each annotator labeled the same tweet as being positive, negative, or neutral. This was based on their impression of the tweet and what sentiment they felt it expressed.

The authors provided the annotators with guidelines regarding the annotation process to help them classify the dialects and sentiments clearly and easily. These depended on the conditions for each label. The annotators followed these guidelines when they classified the dialects:

1) The dialects were classified according to the way that the words were written, replacing the writing, and pronouncing the letter in another way. For example:

- In the Hijazi dialect, the letter "zal,ذ " changes to "dal, د" and the letter "The,ث " changes to "T,ت ".
- In the Najdi dialect, the letter "CAF,ك ", changes to "TS,تس ".
- In the Eastern dialect, the letter "CAF, ك", changes to "j, ج".

2) The classification of the dialects was based on the most utilized words in a specific dialect and was associated with a specific dialect. For example:

- In the Hijazi dialect, "Dahin, دحين ", means now.
- In Najdi dialect, "Halhin, هالحين", means now.
- In the Eastern dialect, "Waid, وايد", means much.

3) If the tweet was difficult to classify or contained words that belonged to two or more dialects, it was classified as unknown or mixed dialect tweets.

4) Symbols that should be used when classifying the dialects (Hijazi: hj, Najdi: nj, Eastern: ea, White dialect: sa).

The following guidelines were carried out when categorizing the sentiment:

1) A tweet was positive if the opinion clearly indicated praise, joy, happiness, and any happy emojis.

2) A tweet was negative if the opinion clearly indicated defamation, sadness, anger, disgust, or any sad emojis.

3) A tweet was neutral if it was not positive or negative, such as news, supplications, or general speech.

Thereafter, some examples of the annotation process are shown in Fig. 3. Moreover, we noted the opinions the annotators gave regarding the dialect annotation, as shown in Table V. For the sentiment analysis, there were three different polarities of the entire dataset as shown in Table VI.

Furthermore, to get the final annotation for both the dialects and sentiments we gathered the more frequently used labels from the annotators for each tweet by using the Mode equation. When there was conflict regarding the annotation, which occurred in 27 tweets, we resolved it by taking the opinion of the fifth annotator as the second stage at annotation process. In case the final annotation was labeled as SA, which happened in only one tweet. We decided to eliminate this tweet and the number of SDCT becomes 4180.

After the percentage of sentiment labels for each dialect had been calculated by observing the annotation process on the corpus. We found that the polarity of neutral dominated in the Saudi tweets, followed by negative and, lastly, positive as shown in Fig. 4. We believe the coronavirus situation is the reason why there was an increase in negative polarity, compared to positive, as we observed that a lot of tweets were related to coronavirus during our study of the tweets.

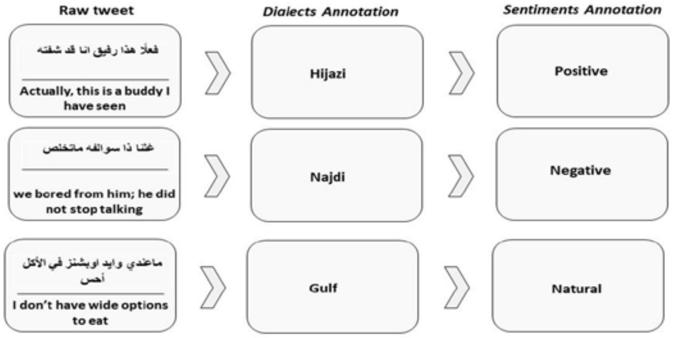
Raw tweet
Dialects Annotation
Sentiments Annotation
فعلًا هذا رفيق انا قد شفته
Actually, this is a buddy I have seen
Hijazi
Positive
غثنا ذا سوالفه ماتخلص
we bored from him; he did not stop talking
Najdi
Negative
ماعندي وايد اوبشنز في الأكل أحس
I don't have wide options to eat
Gulf
Natural

Fig. 3. Examples of Annotations.

TABLE V. ANNOTATION FOR DIALECT LABELS

Label	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Annotator 4**	**Final annotation**
Hijazi	1399	1475	1460	1399	1506
Najdi	1369	1353	1351	1341	1341
Eastern	1333	1325	1334	1337	1333
SA	80	28	36	104	1
Total	4181	4181	4181	4181	4181

TABLE VI. ANNOTATION FOR THE SENTIMENT LABELS

Label	**Annotator 1**	**Annotator 2**	**Annotator 3**	**Annotator 4**	**Final annotation**
Negative	913	921	890	898	925
Positive	787	794	761	770	771
Neutral	2481	2466	2530	2513	2485
Total	4181	4181	4181	4181	4181

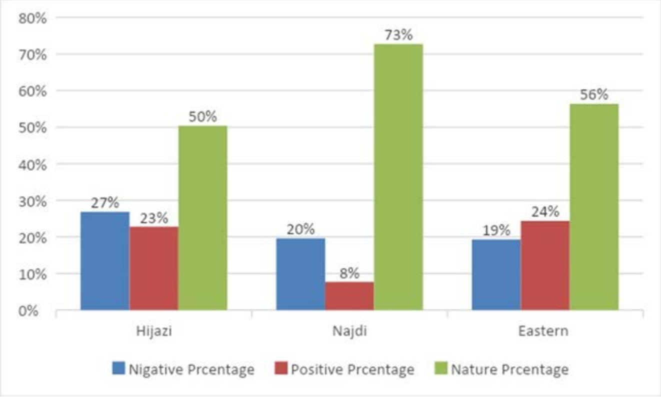

Fig. 4. The Sentiment Labels for each Dialect.

We extracted two hundred words that had been most frequently used in each dialect in the SDTC corpus. This was done graphically as a word cloud of dialect, as shown in Fig. 5, 6 and 7. The font size indicates how frequently the word was used in whole tweets of a specific dialect [24]. Fig. 5 presents the most frequent words in Hijazi Dialect in a word of cloud style. For example, the word "Dahin, دحين "means now and the word "Caman, كمان " means also, they have bigger font in the word cloud that indicates more frequents appearance in the corpus. While, Fig. 6 presents the Najdi Dialect, the word "Ayya, عيا"means refuse, is very common use. From Fig. 7, the Eastern Dialect word cloud shows that the word "Wayed, وايد "means a lot with bigger font to show the frequent used of it.

To ensure the reliability of the results the annotators agreement had to be validated. If the annotators allocated similar labels, we concluded that they all comprehended the annotation instructions in a similar way and would be consistent in their results. The reliability assessments were done to evaluate the level of trust there was in the guidelines and annotation schemes. We used Kappa to calculate the inter-annotator agreements between the four annotators in order to evaluate the quality of the annotations. We calculated the Kappa coefficient for 4,181 tweets that were annotated by the four annotators. The results showed that Kappa was obtained at 0.9382 for the dialect labels, which could be considered almost perfect agreement. The result for the sentiment labels was 0.3199, which indicated fair agreement. Calculation of Kappa coefficient and its interpretation were based on the well-known reliability equations and measurements presented in [25].

D. Validation Test

In this work, two different ways of classifying the dialect were conducted. The aim was to use them to evaluate the accuracy of the SDCT corpus. The first method used all the collected tweets that classified the dialects as Hijazi, Najdi, or

Eastern. The second method established three sub-datasets that had been extracted from the SDCT corpus. Each sub-dataset handled a specific dialect in order to train the model on the characteristics of a particular dialect. Hence, the three sub-datasets were named Hijazi-Dataset, Najdi-Dataset and Eastern-Dataset. The tweets in each sub-dataset were classified as the dialect-name and not-dialect-name. The Hijazi-Dataset contained all the Hijazi tweets in the SDCT corpus that were classified as Hijazi, and Not-Hijazi tweets that had an equal balance of Najdi and Eastern tweets. There was a similar approach with the Najdi-Dataset, which included all the Najdi tweets and an equivalent number of Not-Najdi Tweets. While the Eastern-Dataset was classified as Eastern tweets and Not-Eastern tweets. Table VII illustrates the number of tweets in each dataset.

Fig. 5. Word Cloud of Hijazi Dialect.

Fig. 6. Word Cloud of Najdi Dialect.

Fig. 7. Word Cloud of Eastern Dialect.

TABLE VII. DATASETS OF THE CLASSIFICATION TEST

Dataset	Hijazi	Najdi	Eastern	Not-Dialect	Total
SDCT	1506	1341	1333	-	4180
Hijazi-Dataset	1506	700	700	Not-Hijazi = 1400	2906
Najdi-Dataset	600	1341	600	Not-Najdi = 1200	2541
Eastern-Dataset	650	650	1333	Not-Eastern = 1300	2633

To extract the features, we combined the multi-configurations of the weighting schemes (TF-IDF) and extraction schemes (n-gram), where we tested unigrams, bigrams, trigrams with TF-IDF to present the features vectors. By using these features we trained the most popular classification models, which were named as follows: Linear Support Vector Machines (LSVM), Radial Basis Function Support Vector Machines (RBF SVM), k-Nearest-Neighbors (K-NN), Naive Bayes (NB), Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), AdaBoost, Decision Tree (DT), Bernoulli Naive Bayes (BNB), and Stochastic Gradient Classifier (SGC).

This work employed the k-fold cross-validation technique for training classifier models on each of the four datasets. Cross-validation was used to avoid cases of over-fitting and under-fitting and get a better prediction accuracy. K-folds split the dataset randomly to *k* of splitting. This was done by performing loops through the entire data to apply the classification model and getting an accuracy average. Hence, we proposed using two k-folds, 5-folds and 10-folds, that were best practices. Therefore, we performed eight experiments, including two experiments performed on each dataset. Using the first method of dialect classification, we implemented two main experiments (5-folds and 10-folds) on the SDCT dataset. Table VIII shows the experiments' accuracy. In general, we noticed that the accuracy of 10-folds was better than 5-folds. The highest recorded accuracy of 5-folds in the BNB model was 57.36%. This was also the closest result with DT, GB with bigram and trigram, and SGC with unigram. The worst result was shown to be the k-NN models. For 10-folds, the best result was 69.73% of the SGC model with bigram.

TABLE VIII. THE RESULTS OF THE CLASSIFICATION MODELS IN THE SDCT DATASET

	5-Folds			10-Folds		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	55.36	55.4	55.4	68.82	68.87	68.87
RBF SVM	53.75	53.73	53.73	67.53	67.48	67.48
k-NN	41.5	41.57	41.57	50.69	50.69	50.69
NB	49.33	49.33	49.33	62.77	62.77	62.77
LR	55.52	55.52	55.52	69.04	69.06	69.06
GB	56.94	57.05	57.05	69.35	69.3	69.3
RF	58.39	56.7	56.94	69.28	70.26	70.09
AdaBoost	55.02	53.85	53.92	65.23	65.4	65.4
DT	57.13	57.25	57.15	69.13	69.56	69.56
BNB	57.36	57.36	57.36	68.63	68.68	68.68
SGC	57.03	56.39	56.94	69.57	69.73	69.33
Average	54.30	54.01	54.08	66.37	66.52	66.47

The remaining experiments were performed on the three datasets that had been constructed for the second method of classification. Table IX shows the effect of n-gram with 11 different models on the Hijazi-Dataset with the application of cross-validation. As we can see, AdaBoost had the highest accuracy in both 5-folds and 10-folds. Regarding the Najdi-Dataset, AdaBoost with n-gram = 2,3 achieved the best accuracy of 61.77%. This was done by applying 5-folds. The DT achieved a good result with 77.91% of bigram and a 10-folds configuration, as shown in Table X. Table XI shows the accuracy of the Eastern -Dataset, which had an excellent accuracy of 90% in the SGC model with 5-folds. Regarding 10-folds, the SGC and LSVM resulted in a perfect accuracy of 91%.

TABLE IX. THE RESULTS OF THE CLASSIFICATION MODELS IN THE HIJAZI DATASET

	5-Folds			**10-Folds**		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	71.11	71 .18	71.18	80.21	80.14	80.14
RBF SVM	68.74	68.6	68.6	78.53	78.22	78.22
k-NN	55.03	54.75	54.75	65.4	65.4	65.4
NB	68.6	68.53	68.53	78.02	78.19	78.19
LR	70.32	70.25	70.25	80.39	80.25	80.25
GB	69.57	70.39	70.39	81.73	82.12	82.12
RF	69.91	70.29	68.29	82.48	82.03	83.44
AdaBoost	72.46	72.73	72.73	84.72	84.61	84.61
DT	67.12	67.36	67.67	82.41	82.72	82.52
BNB	66.36	65.98	65.98	83.2	83.07	83.07
SGC	72.04	72.42	72.15	81.35	81.14	81.28
Average	68.29	68.40	68.22	79.85	79.80	79.92

TABLE X. THE RESULTS OF THE CLASSIFICATION MODELS IN THE NAJDI DATASET

	5-Folds			**10-Folds**		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	53.62	53.46	53.46	69.76	69.88	69.88
RBF SVM	53.27	53.19	53.19	72.28	72.17	72.17
k-NN	56.93	56.54	56.54	64.09	63.78	63.78
NB	57.4	57.17	57.17	68.62	68.58	68.58
LR	55.47	55.43	55.43	71.38	71.42	71.42
GB	56.61	57.4	57.4	68.54	68.5	68.5
RF	59.37	60.24	59.33	74.13	74.49	75.51
AdaBoost	61.22	61.77	61.77	74.29	74.53	74.53
DT	60.87	60.39	60.79	77.28	77.91	77.81
BNB	59.84	59.45	59.45	72.01	72.24	72.24
SGC	56.85	57.01	57.52	73.7	74.02	74.41
Average	57.40	57.45	57.45	71.46	71.59	71.71

TABLE XI. THE RESULTS OF THE CLASSIFICATION MODELS IN THE EASTERN DATASET

	5-Folds			**10-Folds**		
Model Name	***unigram***	***bigram***	***trigram***	***unigram***	***bigram***	***trigram***
LSVM	88.48	88.52	88.52	91.36	91.36	91.36
RBF SVM	78.06	78.02	78.02	87.84	87.88	87.88
k-NN	60.09	60.16	60.16	61.67	61.74	61.74
NB	63.05	63.28	63.28	69.45	69.71	69.71
LR	74.96	74.81	74.81	86.66	86.7	86.7
GB	89.58	89.58	89.58	90.64	90.64	90.64
RF	89.24	89.5	89.51	89.81	89.92	89.96
AdaBoost	88.03	87.38	87.38	87.72	81.09	87.69
DT	89.96	90.19	90.36	89.47	89.28	89.39
BNB	73.48	74.61	74.61	81.17	81.09	81.09
SGC	90.34	90.68	90.53	91.44	91.48	91.29
Average	80.47	80.61	80.60	84.29	83.71	84.31

In summary, as we mentioned previously, the 10-folds provided greater accuracy than 5-folds across all eight of the experiments. Furthermore, we observed that, on average, the results of unigram, bigram and trigram were close to each other, particularly in short texts such as tweets. The Eastern-Dataset achieved the best result in this paper compared to the other three datasets with an excellent accuracy of 91.48%.

IV. RESEARCH CHALLENGES

Due to the complex nature of the Arabic language, more investigation is needed, especially in the text mining tools that support the Arabic language. We have encountered a number of obstacles and challenges that need to be taken into account in future works. Some of the obstacles that we faced through the different phases of this study were as follows:

- Data Collected: collecting the tweets that were associated with a particular Saudi dialect was not an easy phase, as most Saudis use the general dialect. In addition, the terms used had significant similarities. Furthermore, the Tweepy library has a limitation when tweets older than one week are retrieved. Hence, a massive effort is required to find the most used unique terms in each region. In addition, a lot of the tweets are advertisements and so it takes a long time in the first cleaning phase to filter them manually.

- Data Preprocessing: There is a limited number of libraries that specialize in Arabic text normalization. This area needs to be highlighted and developed in future works. We believe an excellent specialized library could be an important contribution to facilitating the preprocessing phase.

- Lemmatization: Arabic lemmatization libraries and tools need to be improved as there are insufficient libraries that handle the words in Arabic dialects. As mentioned previously in section III, some lemma roots are completely different, and this impacts their overall meaning. In addition, some libraries don't provide a Python version, which is the most used language in Machine Learning ML.

- Annotation: The similarities between the different dialects means the annotators found it difficult to label some tweets as being in a specific dialect. During the sentiment analysis, some of the tweets were difficult to annotate as being either positive or negative when there was some ambiguity.

V. Conclusion

The objective of this paper was to enrich Arabic, particularly the language used in Saudi Arabia, by constructing a Saudi corpus based on dialects, and make it available for further research in Arabic studies such as NLP applications. This paper presented the methodology used to collect and build a corpus of 4180 multi-dialectal Saudi tweets (SDCT). The corpus was collected by using different keywords, hashtags, and time zones. It was manually annotated into Saudi dialects as Hijazi, Najdi, and Eastern and sentiment as positive, negative, and natural by five native speakers using specific explained guidelines. Cohen's Kappa Coefficient was used to calculate the reliability of the annotations. Eight baseline experiments were performed by using different classifier models with various features and configuration vectors. Four datasets of the corpus were established to fulfill the evaluation of the SDCT corpus that employed a cross-validation mechanism. Further work will be carried out to expand the corpus using other sources for Saudi dialects as well as improve the accuracy of the experiment including various text-features and other factors.

References

[1] Alshutayri and E. Atwell, Exploring Twitter as a source of an Arabic dialect corpus. International Journal of Computational Linguistics (IJCL), 2017. 8(2): p. 37-44.

[2] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, Arabic natural language processing: An overview. Journal of King Saud University-Computer and Information Sciences, 2019.

[3] N. Al-Twairesh, R. Al-Matham, N. Madi, N. Almugren, A.-H. Al-Aljmi, S. Alshalan, R. Alshalan, N. Alrumayyan, S. Al-Manea, and S. Bawazeer, Suar: Towards building a corpus for the Saudi dialect. Procedia computer science, 2018. 142: p. 72-82.

[4] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. B. Shaban, Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. arXiv preprint arXiv:1906.01830, 2019.

[5] S. Harrat, K. Meftouh, and K. Smaïli. Creating parallel Arabic dialect corpus: pitfalls to avoid. 2017.

[6] D. Alahmadi, A. Babour, K. Saeedi, and A. Visvizi, Ensuring Inclusion and Diversity in Research and Research Output: A Case for a Language-Sensitive NLP Crowdsourcing Platform. Applied Sciences, 2020. 10(18): p. 6216.

[7] A. B. Boot, E. T. K. Sang, K. Dijkstra, and R. A. Zwaan, How character limit affects language usage in tweets. Palgrave Communications, 2019. 5(1): p. 1-13.

[8] M. Alruily, Issues of dialectal saudi twitter corpus. Int. Arab J. Inf. Technol., 2020. 17(3): p. 367-374.

[9] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman. Subjectivity and sentiment analysis of Arabic: trends and challenges. in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA). 2014. IEEE.

[10] R. M. Duwairi, N. A. Ahmed, and S. Y. Al-Rifai, Detecting sentiment embedded in Arabic social media–a lexicon-based approach. Journal of Intelligent & Fuzzy Systems, 2015. 29(1): p. 107-117.

[11] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat. Sentiment analysis in arabic tweets. in 2014 5th International Conference on Information and Communication Systems (ICICS). 2014. IEEE.

[12] E. Refaee and V. Rieser. An arabic twitter corpus for subjectivity and sentiment analysis. in LREC. 2014.

[13] M. Nabil, M. Aly, and A. Atiya. Astd: Arabic sentiment tweets dataset. in Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.

[14] A. Assiri, A. Emam, and H. Al-Dossari, Saudi twitter corpus for sentiment analysis. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2016. 10(2): p. 272-275.

[15] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. Procedia Computer Science, 2017. 117: p. 63-72.

[16] A. Alshutayri and E. Atwell. Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers. in OSACT 3 Proceedings. 2018. LREC.

[17] H. Mubarak and K. Darwish. Using Twitter to collect a multi-dialectal corpus of Arabic. in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). 2014.

[18] M. Altamimi and W. J. Teahan, Arabic Dialect Identification of Twitter Text Using PPM Compression. International Journal of Computational Linguistics (IJCL), 2019. 10(4): p. 47-59.

[19] M. Maghfour and A. Elouardighi. Standard and dialectal Arabic text classification for sentiment analysis. in International Conference on Model and Data Engineering. 2018. Springer.

[20] S. Mdhaffar, F. Bougares, Y. Esteve, and L. Hadrich-Belguith. Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. 2017.

[21] Tashaphyne Arabic Light Stemmer. Available online: https://pypi.org/ project/ Tashaphyne/. Accessed: Oct. 5, 2020.

[22] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak. Farasa: A fast and furious segmenter for arabic. in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations. 2016.

[23] M. Altamimi, O. Alruwaili, and W. J. Teahan. BTAC: A Twitter Corpus for Arabic Dialect Identification. in of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018). 2018.

[24] A. L. Uitdenbogerd, World cloud: A prototype data choralification of text documents. Journal of New Music Research, 2019. 48(3): p. 253-263.

[25] J. L. Fleiss and J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 1973. 33(3): p. 613-619.