MSTD: Moroccan Sentiment Twitter Dataset

Soukaina MIHI[1], Brahim AIT BEN ALI[2], Ismail EL BAZI[3], Sara AREZKI[4], Nabil LAACHFOUBI[5]
University Hassan first of Settat Morocco[1, 2, 4, 5]
University Moulay Slimane of Beni Mellal[3]

***Abstract*—With the proliferation of social media and Internet accessibility, a massive amount of data has been produced. In most cases, the textual data available through the web comes mainly from people expressing their views in informal words. The Arabic language is one of the hardest Semitic languages to deal with because of its complex morphology. In this paper, a new contribution to the Arabic resources is presented as a large Moroccan dataset retrieved from Twitter and carefully annotated by native speakers. For the best of our knowledge, this dataset is the largest Moroccan dataset for sentiment analysis. It is distinguished by its size, its quality given by the commitment of annotators, and its accessibility for the research community. Furthermore, the MSTD (Moroccan Sentiment Twitter Dataset) is benchmarked through experiments carried out for 4-way classification as well as polarity classification (positive, negative). Various machine-learning algorithms are combined to feature extraction techniques to reach optimal settings. This work also presents the effect of stemming and lemmatization on the improvement of the obtained accuracies.**

***Keywords*—Sentiment analysis; Moroccan dialect; machine-learning; stemming; lemmatization; feature extraction**

I. Introduction

Natural language processing (NLP) is a very active area of research that exploits the most advanced algorithms and techniques to give machines the ability to understand human language. This branch of Artificial Intelligence has several applications, including translation applications such as Google translate, personal assistance applications such as Cortana, topic detection, sentiment analysis, and others.

Sentiment analysis, also known as opinion mining, is a subfield of NLP that has experienced a strong interest during the last few years. Its study has become inevitable for many businesses wanting to analyze public opinions on the internet. It would be almost impossible for businesses to grow without being able to monitor their presence and brand image through customer interactions.

The applications of sentiment analysis are diverse and closely affecting our daily lives and decisions. In the healthcare field [1][2][3], the opinions published on health communities significantly help patients to find the right doctor for their cases, make the correspondence with their symptoms and take preventive measures. Also, doctors could adapt their prescriptions, schedules, and practices, even pharmaceutical companies benefit by analyzing the public effect of medications and planning studies based on patient attitudes. Moreover, politicians use posts and comments on social media and news articles to determine people's political orientation [4], predict election results [5], and gauge public opinion about changes in legislation or policy projects. In conjunction with the expansion of digital marketing [6], business entities invest in the study of customer perceptions and preferences [7], by analyzing shared opinions about their products and services. Thus, marketers can monitor their brand image and e-reputation. Also, sentiment analysis allows corporations to meet customer expectations and increase their competitiveness.

The expansion of sentiment analysis is made possible, owing to the abundance of data available on social networks. Indeed, several techniques [8] have been proposed to analyze these unstructured data in different languages.

Arab data on social media has enormously increased in recent years. Arab internet users have more access to the internet, which they use every day to get news, share their ideas, buy products online, etc. There are different formats of the Arabic language, including Modern Standard Arabic and Dialectal Arabic. However, when talking about social media, it often implies colloquial forms of expression, users of these platforms create a virtual network of friends with whom they communicate in Dialectal Arabic. This generates more impact and reaches more people.

According to the digital report 2020 for Morocco[1], Internet users represent a percentage of 69% of the Moroccan population with an annual growth of 13%. Today, there are 18 million active users of social networks, with a growth rate of 11% compared to 2019. However, there is very little research that focuses on Moroccan dialectal Arabic, let alone resources available to researchers in sentiment analysis. To the best of our knowledge, there is no publicly available Moroccan dataset for sentiment analysis task. Moreover, it is the need for such a resource to carry out experimentations that led us to develop a large-scale, multi-domain sentiment dataset in Moroccan dialectal Arabic.

The remainder of this paper is organized as follow: Section II details the challenges of analyzing Moroccan dialect as well as twitter posts. Section III presents a survey of works related to the dataset constitution for different dialects, and points out their availability for research community. The next section explains the process of collection and annotation of the MSTD (Moroccan Sentiment Twitter Dataset) to report afterward in Section V the experimentations and results. Finally, we conclude in Section VI.

[1] https://datareportal.com/reports/digital-2020-morocco

II. Challenges

A. Moroccan Dialect Challenges

The Moroccan dialect, widely known as Darija is a variety of Arabic language; it is used in daily communication by Moroccan citizens, Media programs, brand pages on social media, commercial or government advertising to reach out to the general public. Darija is a part of the group of Maghrebi dialects spoken in North African countries and differs itself from one region to another [9]. For example, a distinction is made between the northern dialect, southeast dialect, southwest dialect, and the central dialect, which is the most widespread form.

Moroccan dialect (MD) shares some vocabulary and morphological properties with Modern Standard Arabic (MSA), and is characterized by its own spelling, syntax, lexicon, and phonology. Recently, it has become common to use Darija in writing thanks to the advent of the World Wide Web, blogs, and social media. The following are some of the difficulties in processing Darija in its textual format:

- Code-Switching: The history of Morocco is marked by the French-Spanish colonization, which fed Darija with French and Spanish terms along with the Berber language that is spoken by nearly 40% of the Moroccan population [10]. Consequently, when writing Darija, one can find a mix of MSA, Berber, French, Spanish, and English. An example of this: *عندنا الزلط والشوماج وعر من وباء كورونا بزاف (poverty and unemployment are more dangerous than the pandemic COVID-19)*, here الشوماج is a French word (**chômage/**Unemployment) and (pandemic) **وباء** is a word of MSA.
- Morphological Characteristics: Arabic is a Semitic language characterized by its morphological complexity [11]. It is written from right to left and does not contain capital letters, unlike English. Typically, Darija is very inflectional and derivational. By adding affixes to a word, you get several different words in categories and meanings. Thus, as an example from عجب/Ajaba we can get other words عجيب/Fabulous or عجبني/I like it.
- Orthography: Darija has no orthographic standard [12]. It can be written in the Arabic alphabet (28 letters), Latin (Arabizi). Commonly, on social media, we often find a mix of the two with the use of numbers to write letters that do not exist in the Roman alphabet, such as 3 for ع. (Darija)عجبني/3jebni (Arabizi)/I like it.
- Lack of resources: Research dedicated to Arabic Sentiment Analysis is recent and still scarce compared to other languages. Thus, lexicons and resources for Arabic and particularly dialectal Arabic, are very limited [13].

B. Twitter Challenges

Twitter is a microblogging site that allows users to create personal accounts to share their ideas and activities with followers [14]. Tweets are limited to 140 characters and are usually written in a non-standard format. In Morocco, 17% of Internet users have active accounts on Twitter; they continuously generate massive data which is challenging to process due to:

- Spelling errors: In order to overcome the restrictive length of 140 characters, users may concatenate two or more successive words, delete some letters from long words, and use Acronyms or Slang. Also, a common error is to repeat letters to emphasize a sentiment: عجيييب هادشي/This is fabulous.
- Emoticons: used within a tweet to express a feeling, it is usually the combination of special characters and punctuation. For example, the combination of ":" and ")" usually implies a positive attitude.
- Twitter's features: Twitter is a powerful tool to disseminate information. It offers certain functions for users allowing them to mention other users by adding the "@" character. They can share interesting tweets by using the retweet function "RT". Further, tweets related to the same topic contain the hashtag "#". These special characters create noise in tweets and make the task of processing text from Twitter more difficult than other texts.

C. Annotation Challenges

Annotating textual data for sentiment analysis is not a straightforward process, especially short tweets written in informal language. In the granular level of words, the task of labeling a word is often reduced to the emotion it conveys, is it a positive emotion (love, peace, excitement, etc.) or a negative one (hate, anger, fear, etc.) otherwise, the word is neutral. But as soon as we scale up to the sentence, it becomes complicated depending on the form of the sentence, and the combination of employed tokens [15].

Researchers distinguish three methods of annotation, manual annotation [16], which is based on the intuition of native speakers guided by an instruction scheme and possibly a word lexicon for each target category, automatic labeling [17] that considers the emojis in tweets with reference to an emoticon dictionary for correspondence with the equivalent sentiment and a semantic scoring of the text, and the hybrid method combining both human annotation and automatic techniques [18].

The automatic method can be efficient when managing simple annotation scenarios. It is the categorization of the text [19] in two dimensions of sentiments (positive/negative). However, once we move to a larger scale that concerns several multi-class objectives, with cases of sarcasm sentences, rhetorical questions, mixed sentiments, it is more appropriate to choose manual annotation framed by an annotation scheme and clear instructions.

III. Related Work

Until recently, building resources for sentiment analysis and subjectivity detection was more prevalent for English. These last five years have witnessed a considerable increase of researches devoted to the Arabic language, more specifically, one can find in the literature some recent works regarding the elaboration of materials for the analysis and evaluation of sentiments.

Following sections present various datasets and corpora produced by different research communities within the scope of work on sentiment analysis, with a distinction between three kinds of works, specifically ones related to the MSA, the datasets produced in Vernacular Arabic, and finally resources built on the Maghreb dialects, with an emphasis on research conducted on Tunisian, Algerian, as well as on the Moroccan colloquial languages.

A. MSA Datasets

The most popular datasets in Modern Standard Arabic (MSA) are OCA[20] and AWATIF[21]. The first one was published in 2011. Reviews have been manually extracted directly from various movie websites. Overall, there are 500 reviews from OCA, out of which 250 are positive and 250 negative. Manual processing for cleaning up the text has been performed while the Rapidminer software was used for assessing both Khoja and light stemming.

On the other side, AWATIF is a multi-genre corpus retrieved from three different sources, including Penn Arabic Treebank (PATB) with 2855 sentences, as well as Wikipedia Talk Pages containing 1508 sentences and Web Forums with 1019 sentences. Researchers have cleaned up the noise and text written in dialects to keep only the MSA, and have adopted a manual annotation process in order to compare labeling derived through native speakers with labeling obtained from the Amazon Mechanical Turk crowd-sourcing system. Annotators employed both linguistically motivated and nuance-genre guidelines.

B. Vernacular Arabic Datasets

A considerable effort has been made to build datasets for several Arabic dialects, in particular, the literature is marked by important works concerning the Egyptian dialect. In 2012, Abdul-Mageed and al. reported a corpus consisting of four datasets (DAR-TGRD-THR-MONT) [22] respectively gathered on Maktoob, Twitter, Wikipedia Talk Pages, and Web Forums. In 2013, Mohamed Aly and al. introduced the LABR dataset [23]containing a set of 63K auto-labeled book reviews, employing ratings (1,2) for negative, (4,5) for positive, and 3 for neutral. They eliminated the neutral category and made public the rest of the dataset. Following the same paradigm, Elnagar and al. presented two large scale datasets BRAD [24] for book reviews and HARD[25] for hotel reviews. The ASTD [26] is a popular dataset used extensively for benchmarking sentiment analysis methodologies. 10006 Tweets are covering the following four categories (799 positive tweets, 1684 negative tweets, 6691 objective tweets, 832 mixed tweets) while a manual annotation process was used for labeling all data.

In the same context and sharing the same objective, that of enriching the resources available for sentiment analysis applications in dialectal Arabic, other works have presented datasets in different dialects, including, but not limited to, the following: In Saudi dialect, the Arasenti-tweet [27], a dataset retrieved on Twitter and manually annotated in four classes (positive, negative, neutral and mixed), other datasets have been reported in [28][29]. With regard to the Jordanian dialect, different datasets gathered on Facebook as well as Twitter, were introduced in [30][31][32]. Regarding the Levantine dialect, the works [33][34] yields valuable datasets. Last but not least, a strong focus has been placed on the Sudanese dialect, from which the followings resources can be cited [35][36].

C. Maghrebian Datasets

Maghrebian Arabic is a variety of the vernacular Arabic that is spoken in the North of Africa, including Morocco, Algeria, Libya, Tunisia, and Mauritania. Increasing efforts have been put in processing Maghrebian dialects. In 2018, Rehab and al. presented two Algerian datasets SIAAC[37] & SANA [38], for sentiment polarity identification on newspaper comments. The final corpus consists of 513 manually annotated comments in positive, negative, and neutral classes. They conducted some experiments in which they conclude that KNN outperforms SVM. The work in [39] describes an automated technique used to annotate 8000 Algerian messages into positive and negative, whilst [40] presented a corpus of 10K Facebook comments manually annotated into the two categories positive and negative.

Mdhafar et al. Introduced TSAC [41], a Tunisian corpus collected from Facebook comments and manually annotated, it contains 17K positive and negative comments. Another automated process is presented in [42], which used the Twitter API to collect about 6 million tweets, of which more than 170K written in Maghrebian. For the purpose of validating their approach, the authors have manually tagged 1000 tweets and reported the error rate. The resultant TEAD dataset is the largest Arabic corpus that we know so far. Concerning Libyan Arabic, [43] has recently set up a manual dataset of 2938 tweets annotated in three categories: positive, negative and neutral.

Furthermore, [44] presented a method for automated retrieval from Moroccan tweets according to the geographical localization and trained a Naives Bayes classifier for the multilingual collected dataset. Elouardighi et al. [45] produced a Facebook dataset containing approximately 10K positive and negative comments written in Moroccan and Modern Standard Arabic. Not long ago, [46] investigated deep learning models in processing Moroccan tweets, the authors introduced MSAC, a multi-domain balanced dataset of 2000 positive and negative tweets.

Above, Table I is summarizing the cited datasets that are also a compilation from the most commonly known datasets within the research community, covering the various dialects of Arabic along with their availability to the wider public. The table shows that despite recent efforts to build up resources in Arabic, especially the Moroccan dialect, datasets are not available to carry out studies and benchmark new approaches.

TABLE I. ARABIC DIALECTAL DATASETS FOR SENTIMENT ANALYSIS

Dataset	Size	Type	Classes	Source	Year	Publicly Available
AWATIF[21]	4932	MSA	POS/NEG/NEU/OBJ	WTP/WF/ATB1V3	2012	No
OCA[20]	500	MSA	POS/NEG	Movies Websites	2011	Yes
LABR[23]	63000	Egyptian	POS/NEG/NEU	Goodreads	2013	Yes
BRAD[24]	510600	Egyptian	POS/NEG/NEU	Goodreads	2016	Yes
HARD[25]	93700	Egyptian	POS/NEG/NEU	Booking.com	2016	Yes
ASTD[26]	10006	Egyptian	POS/NEG/OBJ/MIX	Twitter	2015	Yes
AraSenti-Tweet[27]	17573	Saudi	POS/NEG/NEU	Twitter	2017	No
SDTC[28]	5400	Saudi	POS/NEG/NEU/OBJ/SPAM/NOTSURE	Twitter	2018	No
Saudi Twitter Corpus[29]	4700	Saudi	POS/NEG/NEU	Twitter	2016	No
Atoum &al.[30]	3550	Jordanian	POS/NEG/NEU	Twitter	2019	No
Al-harbi &al.[31]	2500	Jordanian	POS/NEG	JEERAN/Jordan	2019	No
Duwairi &al.[32]	22550	Jordanian	POS/NEG/NEU	Twitter	2015	No
ArSenTD-LEV[33]	4000	Levantine	VPOS/POS/VNEG/NEG/NEU	Twitter	2019	Yes
BBN Syrian dataset[34]	2000	Levantine	POS/NEG/NEU	BBN	2015	Yes
SSA-SDA[35]	5456	Sudanese	POS/NEG/NEU	Twitter	2019	No
Abdelhameed &al.[36]	4625	Sudanese	POS/NEG/NEU	Twitter	2019	No
SANA[37]	178	Algerian	POS/NEG/NEU	Newspapers	2019	No
SentiALG[39]	4000	Algerian	POS/NEG	Facebook	2018	No
DzSentiA[40]	49864	Algerian	POS/NEG	Facebook	2019	Yes
TSAC[41]	17000	Tunisian	POS/NEG	Facebook	2017	Yes
TEAD[42]	6m	Tunisian	POS/NEG	Twitter	2017	Yes
Ramadan &al.[43]	2938	Lybian	POS/NEU/NEG	Twitter	2019	No
El Abdouli &al.[44]	930	Moroccan	POS/NEG	Twitter	2017	No
Elouardighi &al.[45]	10254	Moroccan	POS/NEG	Facebook	2017	No
Oussouss &al.[46]	2000	Moroccan	POS/NEG	Twitter	2019	No

IV. DATASET

The proposed model provides a large-scale dataset spanning several domains, including sports, arts, politics, education, and society. The adopted approach is depicted in Fig. 1. First, starting by collecting the data written in Moroccan dialect, then annotating the material into four different classes, after which data is prepared for experimentation; furthermore, the effect of stemming on classification outcomes is benchmarked, and an extended list of stop words is presented, this list is enriched with words from Moroccan dialect to closely observe the influence of using stop-words on determining the correct and incorrect predictions of sentiment categories.

A. Data Collection and Annotation

The collected data is provided from Twitter of users located geographically in Morocco and written only in Arabic. As a result, we have obtained approximately 35K tweets related to the domains of sports, arts, politics, education, and other social issues. While analyzing this data, we have noticed that it contained a high noise content, which makes it very difficult to process it automatically. We stored this data in textual formats, ready to be further analyzed.

A preliminary review of the tweets was performed. Many were cut out because of Twitter's limitation on the number of characters, and some only contained a word or two making reference to an URL and/or a picture. Some were containing nothing other than Hashtags. This initial scan was made to identify the most relevant categories from the collected data. We have noted that we cannot trust emoticons since they convey a fuzzy impression, e.g., Emojis that express a good sentiment were often employed for expressing jokes such as sarcasm and irony. Therefore, the agreement was made for manually labeling these tweets according to the same instructions and some sample annotations Table II.

For annotating tweets, the choice was made for the manual annotation method since it is more performant at giving good quality scores. A team of three native speakers was hired for the task. Total number of tweets was split over three so for each annotator to do one part based upon sample file as well as directions, at the end, an author revises all the annotations so that only those compliant with the initial instructions are kept, thus ensuring consistency with the guidelines established at the outset. We suggest a 4-way annotation that consists of the following four classes (positive, negative, sarcasm, objective).

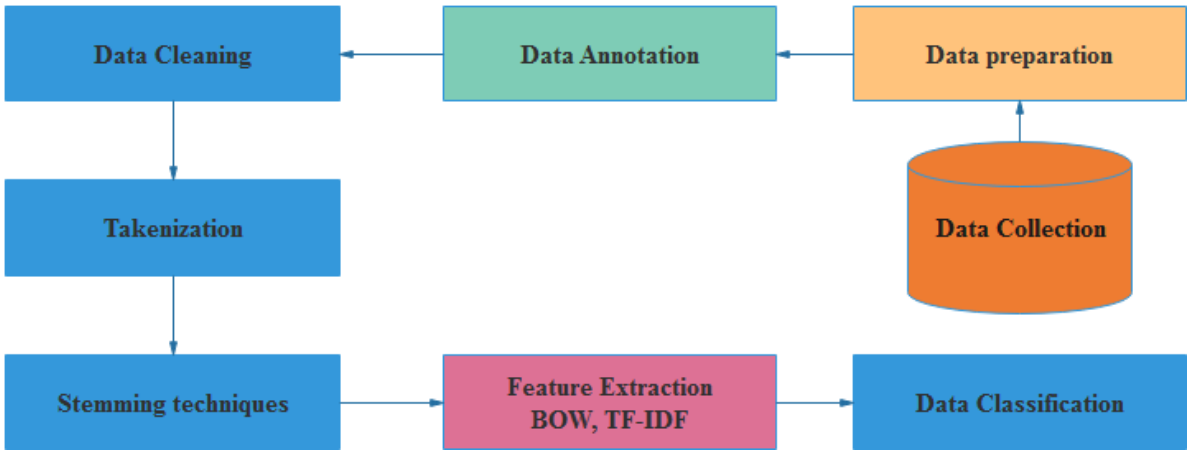

Fig. 1. Proposed System.

Below the guidelines to strive towards identical labeling:

1) Objective: is objectively marked, each sentence reporting a new, although this latter may be written with a negative or positive focus, each fact and description about reality without using expressions and/or qualifiers to convey a sentiment.

2) Positive: we assign a positive category to every statement representing a pleasant emotion from the perspective of the user regarding their own experience, about a public personality, a product or an event, and so forth. Such polarity is characterized by clear terms and the presence of a target of the sentiment.

3) Negative: On the contrary to the positive polarity, we give negative class for tweets that clearly employ some negative words and/or adjectives that express personal beliefs, including colloquial expressions revealing bad perceptions.

4) Sarcasm: A sarcastic text can be characterized either by a contradiction in the statement with the words being used to describe it or by a negative connotation being spoken in a positive way. In addition, there are also some popular sayings or idioms known to Moroccans, which convey sarcasm.

B. Dataset Properties

The overall collected size of the dataset is approximately 35K. All three annotators were given a respective file containing one-third from the tweets, along with instructions and a labeled sampling. It was consented by the guidelines to remove the following tweets:

- Those marked as RT (retweet).
- Tweets in duplicates.
- Tweets that only contain hashtags.
- Tweets that are written entirely in non-Arabic letters.
- Texts representing Spam or insults.
- Tweets with a single word.

The annotators were able to annotate over 12K tweets with four separate classes: 6378 objective, 2769 negative, 866 positives, and 2188 sarcasm, as shown in Fig. 2.

In order to illustrate the frequency of tokens in the MSTD dataset corresponding to a given class, the following word cloud schemes in Fig. 3, Fig. 4, Fig. 5 and Fig. 6 show the word occurrences for the four labeled classes.

TABLE II. ANNOTATED SAMPLE

Text	Translation	Class	Label
البوليس ديالنا احسن بوليس ماتفلت معاه حتى حاجة،واش غلبوكم الشفارة ولاخايفين منهم	"Our police are the best in the world, did the thieves beat you or what ?!"	Sarcasm	3
أجمل مدينة كانت مفخرة جهة سوس العالمية الجميلة	"The most beautiful city, it's our proud, beautiful international Souss."	Positive	2
لحظة اعتقال تلميذ غش في امتحانات الباكالوريا	"The moment when a student stops for the baccalaureate exam."	Objective	0
العلاقة بيناتنا متوترة بسبب الإختلاف في وجهات النظر	"Our relationship is upset because of the difference of opinion."	Negative	1

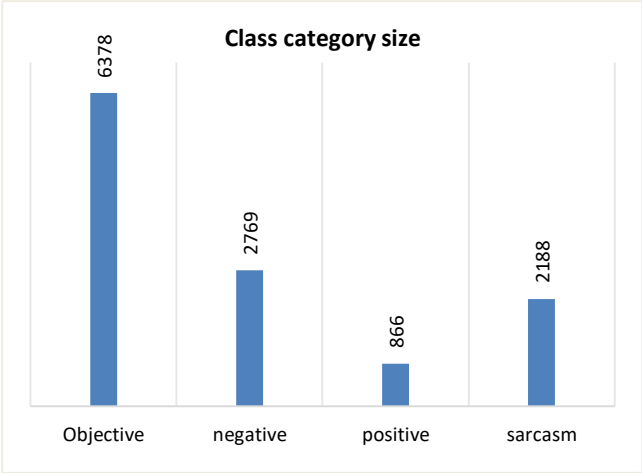

Fig. 2. Class Category Size.

Fig. 3. Word Occurrence for Objective Class.

Fig. 4. Word Occurrence for Negative Class.

Fig. 5. Word Occurrence for Positive Class.

Fig. 6. Word Occurrence for Sarcasm Class.

C. Data Preprocessing

Data cleaning constitutes an extremely essential phase for NLP tasks, most notably for Arabic sentiment analysis, as proven through several studies [47][48]. Processing text in its plain format substantially enhances accuracy measures as well as reduces redundancy and irrelevant tokens to detect corresponding classifications. A cleaning phase of collected tweets was processed after annotating them.

The first step is to remove all special characters, emoticons, Hashtags, URLs, Html code, usernames, punctuation, and non-Arabic letters. This stage enables to reduce noise in text extensively. Then, we normalized the data by removing the diacritics (tashdid, Fatha, Tanwin Fath, Damma, Tanwin Damm, Kasra, Tanwin Kasr, Sukun, Tatwil). These signs that are placed above or under the Arabic letters can alter the meaning of a single word, for instance (مَدْرَسَةٌ means school while مُدَرِّسَةٌ means teacher). The last phase of normalization consisted of normalizing the various letters shapes (ة,ا,ء) based on the Pyarabic library[2].

Besides, stop words are words that do not add any sentiment to text; we have built up a list that contains Arabic and specific Moroccan stop words, which we removed from cleaned tweets to minimize the size of the features. The final step involved retrieving tokens from the text by performing the tokenization. Pyarabic library and NLTK provide tools for doing this tokenization.

D. Data Stemming and Lemmatization

Arabic has a high inflectional degree since several words may be derived out of the same letters, although these words are necessarily synonyms in a sense. Stemming is a process consisting of the reduction of forms from each word into its root by cutting up prefixes and suffixes found in the beginning and the end of words. In contrast, lemmatization is a method with the same purpose to find the basis of the words but takes into account the morphological nature of words to extract a lemma that conserves meaning. Stemming can be classified into two types: root based stemming or heavy stemming and light stemming:

1) Root-based stemming: these methods attempt to use linguistic and heuristic analysis to extract the basic root of a word by removing its longest prefixes and suffixes. Despite its wide use, root-based methods may lead to different semantic meanings for the same extracted root. Example of use: considering the word "ينتصرون/they win", a root-based algorithm will remove "ي" and "ون" resulting in "نتصر" which has no semantic meaning.

2) Light-stemming: on its side, light stemming doesn't use deep linguistic analysis. Instead, it performs some heuristic analysis to strip off most frequent prefixes and suffixes attached to the beginning and the end of words, and it results in fewer errors in semantic meaning. Example of a light-stemming method: from the word "الخدمات/services", "ال" will be removed from the beginning and "ات" from the end. As a result, the root will be "خدم".

3) Lemmatization: uses morphological analysis and relies on vocabulary usage to derive lemma from words. Some studies show that lemmatizers lead to better results than stemmers. Example of a lemmatization method: considering the word "ينتصرون/they win", a lemmatizer algorithm will remove "ي" and "ون" resulting in "انتصر", a word that means "won".

[2] T. Zerrouki, Pyarabic, An Arabic language library for Python, https://pypi.python.org/pypi/pyarabic/, 2010.

V. Experimentation and Results

Within the scope of this research, we have investigated the multi-class classification issue through benchmarking the most popular classifiers for the NLP tasks to conduct the 4-way classification on unbalanced dataset firstly as well as to detect polarity for a balanced configuration. The results are compared for each method by combining the feature extraction methods with the machine learning algorithms. Then we try also to address how stemming impacts the classification of text that is written only in informal Moroccan Arabic with different local dialects.

A. Feature Extraction

Feature extraction represents an essential process for optimizing machine learning models. It allows to avoid the over-fitting problem reducing the dimensionality of data. Thus, a selection is made of the most relevant features, which significantly enhances accuracy as well as decreases the time required in training models.

Two different feature extraction algorithms were used: the bag of words, this model provides a number vector that represents the occurrence frequencies associated with words, and TF-IDF (Term Frequency-Inverse Document Frequency), which does not just simply calculate the word occurrence, but assigns an individual TF-IDF score for each word as a ratio of two measurements: The TF representing a word's frequency within a document and the IDF computing its importance within a dataset, whereby the words appearing least often in documents are those that are most prominent for classification purposes. For both feature extraction models, we experiment the N-gram features.

B. Classification Algorithms

The classification phase is handled through different machine learning algorithms benefiting from the implementation of the python sklearn library. Default settings were adopted to compare the following algorithms: SVC, Bernouli Naïve Bays, Multinomial Naives Bayes, LinearSVC, LogisticRegression, RandomForest, XGBoost, K-Nearest Neighbors, and the DecisionTree.

C. Stemming Techniques

Different experiments were conducted on two stemmers as well as a lemmatizer. For this, the ISRI-Stemmer [49] algorithm is tested for root-based methods, the Tashafyn [3] algorithm for light stemming, and the Farasa [50] framework for lemmatization.

D. Discussion

Table III presents the accuracy obtained by the selected algorithms for both vectorization methods (BOW, TF-IDF) combined with the stemming algorithms (ISRI Stemmer, Tashafyn light Stemmer). In order to evaluate the effect of stemming, also experiments were conducted on the preprocessed dataset without the stemming phase.

From the results achieved for the 4-way classification, it is obvious that the Logistic Regression algorithm outperforms the SVC algorithm. Indeed, the highest accuracy, 0.563, was given from experimentation Logistic Regression + ISRI Stemmer + TF-IDF tri-gram. Whilst using the Tashafyn light stemmer system, it was found out that the most accurate settings are SVC+Tashafyn+ BOW tri-gram with 0.556. We notice that for any algorithm of classification and whatever is the vectorization process, the stemming phase represents an essential and reliable way of improving both performance and accuracy rate.

Table IV shows obtained results using the balanced setup for polarity classification (positive, negative), and we consider the details of accuracy measurements concerning on one side, the ISRI Stemmer algorithm as well as those regarding the Lemmatizer Farasa with Ngram of both BOW Ngram and TF-IDF schemes.

The recorded findings demonstrate similarity in accuracy provided by each of the classification algorithms. With the TF-IDF model, the highest scoring is achieved by using the 1-gram scheme, in contrast with the BOW model, which has higher scores given using the 3-gram scheme. While comparing both methods Farasa and ISRI Stemmer, we can observe that with Farasa lemmatizer, a better overall accuracy is achieved when using SVM (SVC and LinearSVC) and Logistic Regression algorithms. At this point, it is worth mentioning the very fast processing time of the Logistic Regression algorithm with regard to SVC.

[3] https://pypi.org/project/Tashaphyne/

TABLE III. 4-WAY CLASSIFICATION EXPERIMENTATION RESULTS

Stemming Mechanism	Classifier	TF-IDF			Bag of Words		
		1g	1g+2g	1g+2g+3g	1g	1g+2g	1g+2g+3g
ISRI Stemmer	SVC	0.561	**0.562**	0.550	**0.559**	0.555	0.556
	XGB	0.552	0.551	0.544	0.547	0.547	0.547
	BNB	0.555	0.555	0.558	0.558	0.556	0.555
	LogisticRegression	0.557	0.557	**0.563**	0.554	0.552	0.548
	KNN	0.519	0.514	0.517	0.502	0.503	0.502
	DecisionTree	0.476	0.481	0.489	0.476	0.481	0.484
	RandomForest	0.555	0.548	0.544	0.548	0.551	0.548
Tashafyn Light Stemmer	SVC	0.554	0.551	0.550	0.550	0.555	**0.556**
	XGB	0.545	0.545	0.540	0.544	0.548	0.547
	BNB	**0.555**	0.547	0.549	0.553	0.557	0.555
	LogisticRegression	0.548	0.548	0.546	0.539	0.552	0.548
	KNN	0.513	0.519	0.520	0.525	0.503	0.502
	DecisionTree	0.476	0.469	0.486	0.484	0.479	0.482
	RandomForest	0.551	0.542	0.539	0.546	0.553	0.547
No-Stemming	SVC	0.549	0.547	0.547	0.548	0.546	0.546
	XGB	0.541	0.540	0.539	0.539	0.540	0.541
	BNB	0.555	0.546	0.546	0.549	0.549	0.549
	LogisticRegression	0.548	0.554	0.554	0.545	0.549	0.550
	KNN	0.499	0.503	0.503	0.510	0.517	0.519
	DecisionTree	0.479	0.488	0.488	0.479	0.485	0.473
	RandomForest	0.548	0.546	0.548	0.549	0.540	0.538

TABLE IV. POLARITY CLASSIFICATION EXPERIMENTATION RESULTS

		TF-IDF			BOW		
		1g	1g+2g	1g+2g+3g	1g	1g+2g	1g+2g+3g
ISRI Stemmer	SVC	0.756	**0.761**	0.747	**0.716**	0.713	0.713
	BNB	**0.670**	0.606	0.552	0.718	0.739	**0.744**
	LR	0.741	0.744	**0.750**	0.739	0.741	**0.744**
	MNB	0.750	**0.759**	0.750	0.756	0.773	**0.773**
	LinearSVC	**0.747**	0.747	0.736	0.730	0.713	**0.716**
FARASA	SVC	0.767	**0.772**	0.764	0.707	0.710	**0.710**
	BNB	**0.695**	0.606	0.537	0.730	0.747	**0.747**
	LR	**0.772**	0.761	0.753	0.733	0.747	**0.747**
	MNB	**0.756**	0.741	0.736	0.748	0.750	**0.750**
	LinearSVC	0.773	0.767	**0.776**	**0.698**	0.672	0.672

VI. CONCLUSION AND FUTURE WORK

Within this work's scope, a major contribution is presented aiming constitution of resources for sentiment analysis in Arabic language, especially the Moroccan dialect spoken by over 35 million people and widely spread through social media and awareness-raising programs, such as during the Covid-19 pandemic. We have collected and labeled on a large scale-dataset containing over 12k tweets, which is publicly available[4] for the research community. A particularity with this dataset resides in its handling of various classes: positive, negative, sarcasm, and objective. This assumption comes from the fact that most of the posts on twitter emanate from the younger generation who in most cases, express themselves implicitly through expressions of sarcasm. In another part, substantial experiments were performed for the validation of

[4] https://github.com/moroccanSA-NER/SA-Moroccan

the MSTD dataset in both unbalanced and balanced configurations, also in order to compare the accuracy for 4-way classification and the classification of two classes. In addition, the effect of stemming on the improvement of the results was investigated, this leads to the conclusion that for the case of Arabic language, lemmatization remains a reliable choice. The annotation and processing of the dataset is really tedious and time-consuming work. However, we hope through manual annotation to contribute with a precise and accurate dataset.

Among the limitations encountered in developing this work is the computational capacity of the machines available for the experiments. However, the obtained outcomes of this study motivate us further to pursue this work at several levels, including an investigation into deep learning models as well as additional feature extraction methods in order to improve multi-class classification scoring. Furthermore, this dataset could be extended to include more valuable texts coming out of Facebook and/or YouTube.

REFERENCES

[1] S. Gohil, S. Vuik, and A. Darzi, "Sentiment analysis of health care tweets: Review of the methods used," J. Med. Internet Res., vol. 20, no. 4, 2018.

[2] V. L. Mane, S. S. Panicker, and V. B. Patil, "Summarization and sentiment analysis from user health posts," 2015 Int. Conf. Pervasive Comput. Adv. Commun. Technol. Appl. Soc. ICPC 2015, vol. 00, no. c, 2015.

[3] F. C. Yang, A. J. T. Lee, and S. C. Kuo, "Mining Health Social Media with Sentiment Analysis," J. Med. Syst., vol. 40, no. 11, 2016.

[4] A. P. Patil, D. Doshi, and D. Dalsaniya, "Applying Machine Learning Techniques for Sentiment Analysis in the Case Study of Indian Politics." eds) Advances in Signal Processing and Intelligent Recognition Systems. SIRS 2017. Advances in Intelligent Systems and Computing, vol 678. Springer, Cham. https://doi.org/10.1007/978-3-319-67934-1_31.

[5] T. Elghazaly and A. Mahmoud, "Political Sentiment Analysis Using Twitter Data," ICC '16: Proceedings of the International Conference on Internet of things and Cloud Computing.March 2016.

[6] D. Al-Hajjar and A. Z. Syed, "Applying sentiment and emotion analysis on brand tweets for digital marketing," 2015 IEEE Jordan Conf. Appl. Electr. Eng. Comput. Technol. AEECT 2015, 2015.

[7] A. Alamsyah, W. Rahmah, and H. Irawan, "Sentiment analysis based on appraisal theory for marketing intelligence in Indonesia's mobile phone market," J. Theor. Appl. Inf. Technol., vol. 82, no. 2, pp. 335–340, 2015.

[8] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, 2014.

[9] J. Heath, Jewish and muslim dialects of moroccan arabic. 2013.

[10] Y. Samih and W. Maier, "An Arabic-moroccan Darija code-switched corpus," Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016, pp. 4170–4175, 2016.

[11] S. Al-Osaimi and M. Badruddin, "Sentiment Analysis Challenges of Informal Arabic Language," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 2, pp. 278–284, 2017.

[12] R. Tachicart, K. Bouzoubaa, and H. Jaafar, "Lexical differences and similarities between Moroccan dialect and Arabic," Colloq. Inf. Sci. Technol. Cist, pp. 331–337, 2017.

[13] L. Albraheem, "Exploring the problems of Sentiment Analysis in Informal Arabic," IIWAS '12: Proceedings of the 14th International Conference on Information Integration and Web-based Applications & ServicesDecember 2012 Pages 415–418.

[14] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and A. R. Montejo-Ráez, "Sentiment analysis in Twitter," Nat. Lang. Eng., vol. 20, no. 1, pp. 1–28, 2014.

[15] S. Mohammad, "A Practical Guide to Sentiment Annotation: Challenges and Solutions," no. January, pp. 174–179, 2016.

[16] R. F. de Azevedo, J. P. Santos Rodrigues, M. R. da Silva Reis, C. M. C. Moro, and E. C. Paraiso, "Temporal Tagging of Noisy Clinical Texts in Brazilian Portuguese. International Conference on Computational Processing of the Portuguese Language (PROPOR)," Lncs, vol. 11122, pp. 231–241, 2018.

[17] A. Baccouche, B. Garcia-Zapirain, and A. Elmaghraby, "Annotation Technique for Health-Related Tweets Sentiment Analysis," 2018 IEEE Int. Symp. Signal Process. Inf. Technol. ISSPIT 2018, no. December, pp. 382–387, 2019.

[18] L. Y. F. Su, M. A. Cacciatore, X. Liang, D. Brossard, D. A. Scheufele, and M. A. Xenos, "Analyzing public sentiments online: combining human- and computer-based content analysis," Inf. Commun. Soc., vol. 20, no. 3, pp. 406–427, 2017.

[19] C. Bosco, V. Patti, and A. Bolioli, "Developing corpora for sentiment analysis: The case of irony and senti-TUT," IJCAI Int. Jt. Conf. Artif. Intell., vol. 2015-Janua, no. Ijcai, pp. 4158–4162, 2015.

[20] X. Liu, "Full-Text Citation Analysis : A New Method to Enhance," J. Am. Soc. Inf. Sci. Technol., vol. 64, no. July, pp. 1852–1863, 2013.

[21] M. Abdul-Mageed and M. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," Proc. 8th Int. Conf. Lang. Resour. Eval. Lr. 2012, no. April 2015, pp. 3907–3914, 2012.

[22] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," Comput. Speech Lang., vol. 28, no. 1, pp. 20–37, 2014.

[23] M. Aly and A. Atiya, "LABR: A large scale arabic book reviews dataset," ACL 2013 - 51st Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., vol. 2, no. October 2014, pp. 494–498, 2013.

[24] A. Elnagar and O. Einea, "BRAD 1 . 0 : Book Reviews in Arabic Dataset," 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA).

[25] A. Elnagar, Y. S. Khalifa, and A. Einea, "Hotel arabic-reviews dataset construction for sentiment analysis applications," Stud. Comput. Intell., vol. 740, pp. 35–52, 2018.

[26] M. Nabil, M. Aly, and A. Atiya, "ASTD: Arabic Sentiment Tweets Dataset," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, no. September, pp. 2515–2519, 2015.

[27] 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 Comput. Sci., vol. 117, pp. 63–72, 2017.

[28] A. Al-thubaity and M. Alharbi, "A Saudi Dialect Twitter Corpus for Sentiment and Emotion Analysis," 2018 21st Saudi Comput. Soc. Natl. Comput. Conf., pp. 1–6, 2018.

[29] A. Assiri, A. Emam, and H. Al-dossari, "Saudi Twitter Corpus for Sentiment Analysis," International Journal of Computer and Information Engineering Vol:10, No:2, 2016.

[30] J. O. Atoum and M. Nouman, "Sentiment Analysis of Arabic Jordanian Dialect Tweets," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019.

[31] O. Al-harbi, "Classifying Sentiment of Dialectal Arabic Reviews : A Semi-Supervised Approach," International Arab Journal Of Information Technology, 16(6):995:1002.

[32] R. M. Duwairi, "Sentiment Analysis for Dialectical Arabic," 2015 6th International Conference on Information and Communication Systems (ICICS) pp. 166–170, 2015.

[33] 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," Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[34] M. Salameh, "Sentiment after Translation : A Case-Study on Arabic Social Media Posts Sentiment after Translation : A Case-Study on Arabic Social Media Posts," Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies no. June 2016, 2015.

[35] M. E. M. Abo, N. A. K. Shah, V. Balakrishnan, M. Kamal, A. Abdelaziz, and K. Haruna, "SSA-SDA: Subjectivity and sentiment

analysis of sudanese dialect Arabic," 2019 Int. Conf. Comput. Inf. Sci. ICCIS 2019, pp. 1–5, 2019.

[36] H. Al-Rubaiee, R. Qiu, K. Alomar, and D. Li, "Sentiment Analysis of Arabic Tweets in e-Learning," J. Comput. Sci., vol. 12, no. 11, pp. 553–563, 2016.

[37] P. Silhavy, "Applied Computational Intelligence and Mathematical Methods," vol. 662, 2018.

[38] H. Rahab, A. Zitouni, M. Djoudi, and S. Sentiment, "SANA : Sentiment analysis on newspapers comments in Algeria To cite this version : HAL Id : hal-02444686 SANA : Sentiment Analysis on Newspapers comments in Algeria," J. King Saud Univ. - Comput. Inf. Sci., 2020.

[39] I. Guellil, A. Adeel, F. Azouaou, and A. Hussain, SentiALG: Automated Corpus Annotation for Algerian Sentiment Analysis, vol. 10989 LNAI, no. Ml. Springer International Publishing, 2018.

[40] A. Abdelli, F. Guerrouf, O. Tibermacine, and B. Abdelli, "Sentiment Analysis of Arabic Algerian Dialect Using a Supervised Method," Proc. - 2019 Int. Conf. Intell. Syst. Adv. Comput. Sci. ISACS 2019, 2019.

[41] S. Medhaffar, F. Bougares, Y. Estève, and L. Hadrich-Belguith, "Sentiment Analysis of Tunisian Dialects: Linguistic Ressources and Experiments," no. January, pp. 55–61, 2017.

[42] H. Abdellaoui and M. Zrigui, "Using tweets and emojis to build TEAD: An arabic dataset for sentiment analysis," Comput. y Sist., vol. 22, no. 3, pp. 777–786, 2018.

[43] R. Alfared, "A Topic-based Twitter Sentiment Analysis Training Dataset for Libyan Dialect," (IJACSA) International Journal of Advanced Computer Science and Applications, no. March, pp. 4–6, 2019.

[44] A. El Abdouli, L. Hassouni, and H. Anoun, "Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm," International Journal of Computer Science and Information Security, vol. 15, no. 12, 2017.

[45] Maghfour M., Elouardighi A. (2018) Standard and "Dialectal Arabic Text Classification for Sentiment Analysis". In: Abdelwahed E., Bellatreche L., Golfarelli M., Méry D., Ordonez C. (eds) Model and Data Engineering. MEDI 2018. Lecture Notes in Computer Science, vol 11163. Springer, Cham. https://doi.org/10.1007/978-3-030-00856-7_18.

[46] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," J. Inf. Sci., 2019.

[47] R. Duwairi and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text," J. Inf. Sci., vol. 40, no. 4, pp. 501–513, 2014.

[48] M. Sawalha et al., "Enhancing the Arabic sentiment analysis using different preprocessing operators.," New Trends Inf. Technol., no. April, pp. 113–117, 2017.

[49] M. G. Syarief, O. T. Kurahman, A. F. Huda, and W. Darmalaksana, "Improving Arabic Stemmer: ISRI Stemmer," Proceeding 2019 5th Int. Conf. Wirel. Telemat. ICWT 2019, 2019.

[50] H. Mubarak, "Build fast and accurate lemmatization for Arabic," Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval., pp. 1128–1132, 2019.