MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis

Moncef Garouani, Jamal Kharroubi

MAC : An open and free Moroccan Arabic corpus for sentiment analysis

Moncef Garouani[1,2] and Jamal Kharroubi[1]

[1] LISIC Laboratory, Univ. Littoral Cote d'Opale Calais, France
[2] LSIA Laboratory, Faculty of sciences and techniques Fez, USMBA, Morocco
moncef.garouani@etu.univ-littoral.fr,jamal.kharroubi@usmba.ac.ma

Abstract. The proliferation of social media has allowed Internet users to post their views and opinions online. This generated a vast amount of raw data in informal ways. For many organizations and individuals, this data is vital for providing insight into future decisions. Preprocessed corpora are considered as the basic requirement for the development and evaluation of opinion mining (OM) systems. However, the vast majority of corpora intended for OM research are not large and free for the researchers' community. This lack of free and large OM corpora represents a major obstacle for promoting research on sentiment analysis systems, especially for rich and complex languages as the Moroccan Arabic (MA) one. To overcome this gap, this paper presents a new contribution to the MA resources. A free and large Moroccan Arabic corpus consisting of 18000 manually labeled tweets resulting in a lexicon-dictionary of 30000 words labeled as positive, negative and neutral. To the best of our knowledge, MAC (Moroccan Arabic Corpus) is the first open and largest MA corpus for sentiment analysis. It is pioneer by its size, its quality given by the consistency of the native annotators (IAA=0.9), and its accessibility to the research community. The MAC is benchmarked for forthcoming works through an exploratory data analysis carried out using the two-sentiment analysis approaches for polarity classification as well as language identification. In addition, the MAC corpus along with the necessary code to explore it have been released.

Keywords: Sentiment analysis, Opinion mining, Moroccan dialect, Arabic, Data Mining, Machine learning.

1 Introduction

With the worldwide spread use of social media platforms, opinion mining has become a useful and prominent technique to capture public opinions and views in different disciplines. OM or sentiment analysis (SA) refer to a computational process to automatically extract individuals' opinions, feelings, or attitudes towards particular events or issues [1, 2]. SA has a vital function in real-life applications and decision-making processes in various domains for individuals and organisms [2].

Sentiment analysis is a subfield of Natural Language Processing (NLP) that has gained a strong interest during the last few years. However, the detection of Arabic sentiment polarity is a challenging task, due to limitations of data resources and the language complexity, particularly for its vernacular varieties [2]. It has many types and forms, including the Classical Arabic (CA),language found in the Holy Quran, Modern Standard Arabic (MSA) used in education, newspapers and formal speech, and Dialectal Arabic (DA), which is the informal everyday spoken language, used in chat rooms and social media platforms.

In 2021, Morocco reached 20 million social media users with an annual growth of 23% compared to 2019 (Statista[1]). However, very little research focuses on the Moroccan vernacular Arabic due to the lack of free and publicly available corpus and lexicon resources than for other Arabic dialects [2, 3]. For instance, the Levantine Arabic [4], as has the Egyptian dialect have been heavily studied, however, resources used for one Arabic country cannot be applied to another. Thus, there is still a need for Arabic corpora, especially for DA [5].

This research aims to fill this gap, by creating a MA benchmark corpus and a lexicon-dictionary for use in data mining. In this paper we focuse on Moroccan Arabic Sentiment Analysis and we provide solutions to one of the challenges that faces SA by creating the largest MA corpus and lexicon-dictionary. This resource is based on data extracted from Twitter. It also provides an evaluation of this corpus, further demonstrating its quality and applicability.

The remainder of this paper is organized as follows : Section II provides an overview of works related to Arabic data mining, the constituent dataset for different Maghrebian dialects, and points out their availability for the research community. The process of collecting and annotating the MAC is described in Section III, to report afterward in Section IV the benchmarking experimentations and results. Finally, conclusions are drawn in Section V.

2 Related work

2.1 Sentiment analysis approaches

The problem of Sentiment Analysis has been widely studied on the European languages and the Asian ones [**el-garouaniSentimentAnalysisMoroccan2021**]. The proposed solutions are largely dominated by the use of two typical analysis approaches : supervised (corpus-based) and unsupervised (lexicon-based) approach.

Corpus-based approach The Corpus-based approach is a data-driven approach that requires a large labelled data to train a classifier such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes (NB), and K-Nearest Neighbor (KNN).

[1] https://www.statista.com/

Lexicon based approach The Lexicon based approach makes use of the words and their corresponding polarities. In sentiment analysis, the term *lexicon* is a synonym for the word dictionary, however, rather than words and their definitions in dictionaries, lexicons in sentiment analysis contain the words along with their corresponding polarities [6]. That is, every word is associated with its sentiment orientation. The lexicon could be created from existing dictionaries or from a corpus [7]. Table 1 provides a compact summary of recent works on SA for Arabic text.

Table 1: List (sample) of recent works on sentiment analysis.

Reference	Language	Approach	Dataset size	Performance	
				Criteria	Result
[7]	Jordanian	Lexical	900	Accuracy	86.89%
[6]	Jordanian	Lexical	2000	Accuracy	70.05%
[1]	Algerian	supervised	49864	Accuracy	86%
[8]	Algerian	supervised	4000	F1-score	78%
[9]	Moroccan	supervised	9901	F1-score	84.33%
[10]	Moroccan	supervised	2000	Accuracy	93%
[2]	Moroccan	supervised	13550	Accuracy	92.09%

Since the focus of this work is on data resources, the following section survey the corpora produced by different research communities within the scope of work on sentiment analysis, with an emphasis on research conducted on the Arabic language (standard and vernacular) especially the Moroccan one.

2.2 Maghrebian corpora

The Maghrebi Arabic is a variety of vernacular Arabic spoken primarily in the Maghreb region (notably Morocco, Mauritania, Algeria, Libya and Tunisia). An increasing effort is being made to process Maghrebian dialects [4]. In 2018, for sentiment polarity identification from newspaper comments, Rehab et al. presented SANA [11]. An Algerian dataset of 178 comments annotated into positive, negative, and neutral categories. In their evaluation study, they found that the KNN classifier outperformed the SVM. Similarly, in [1], 10K Facebook comments were manually classified into positive and negative polarities, whilst the work in [8] describes an automated technique for categorizing 8000 Algerian messages into positive and negative comments.

Mdhafar et al. presented TSAC [12], a Tunisian corpus of 17K comments collected from the Facebook social media and manually annotated in positive and negative comments. Another automated process, described in [13], gathered about 6 million tweets using the Twitter API, of which more than 170K

were written in Maghrebi Arabic. However, the authors have tagged just 1000 tweets. Concerning Libyan Arabic, the authors of [14] have recently put together a manually annotated corpus of 2938 tweets categorized as positive, negative and neutral.

Not long ago, a method to extract Moroccan tweets by geographical localization has been presented [3]. Besides, [10] introduced ASA, a multi-domain corpus of 2000 Moroccan tweets tagged as positive and negative tweets. Likewise, [9] gathered a dataset of approximately 10K positive and negative Facebook comments written in Moroccan and Modern Standard Arabic. Table 2 shows more details about some of the state-of-the-art Arabic corpora.

Table 2: Arabic corpora for sentiment analysis.

Dataset	Size	Arabic	Classes	Source	Year	Publicly Available
[15]	10006	Egyptian	4	Twitter	2015	✓
[16]	17573	Saudi	3	Twitter	2017	✘
[17]	5400	Saudi	6	Twitter	2018	✘
[12]	17000	Tunisian	2	Facebook	2017	✓
[13]	6m	Tunisian	2	Twitter	2017	✓
[14]	2938	Lybian	3	Twitter	2019	✘
[11]	178	Algerian	3	Newspapers	2019	✘
[1]	49864	Algerian	2	Facebook	2019	✓
[3]	930	Moroccan	2	Twitter	2017	✘
[9]	10254	Moroccan	2	Facebook	2017	✘
[10]	2000	Moroccan	2	Twitter	2019	✓
[4]	12K	Moroccan	4	Twitter	2020	✘

As can be seen, compared to other languages, the Moroccan dialect has no large-scale free available corpora to carry out studies and benchmark new approaches. In addition, some of the public available ones provide no information about the annotation, which may limit their use. This paper aims to fill this gap by presenting the creation and annotation details of a large-scale Moroccan Arabic corpora and lexicon-dictionary. In addition, we will make it freely available to the research community.

3 Benchmark corpus

The constructed Benchmark provides a large-scale corpus spanning several domains, including sports, arts, politics, education, and society. The data collection, filtering, cleaning, and pre-processing steps are illustrated in Figure 1.

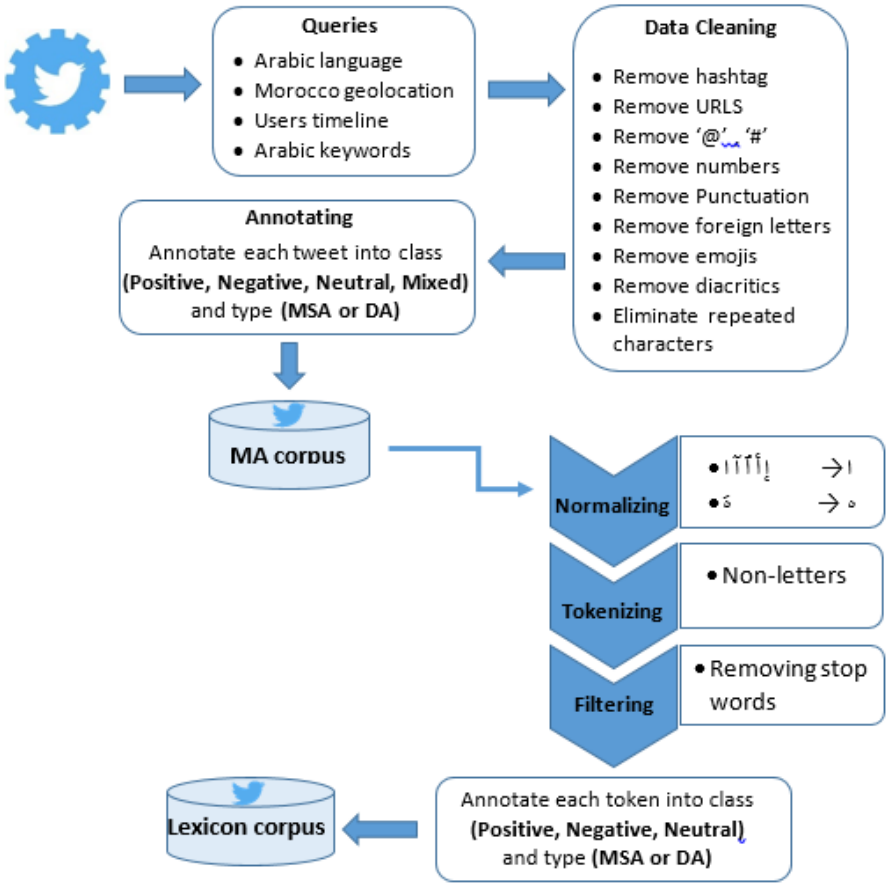

Fig. 1: The flow chart of the proposed construction approach.

3.1 Data collection

To build the MAC corpus, we have developed a Python program based on the Twitter API[2] for the extraction of data. This API provides access to Twitter data (tweets, users informations, etc.). Our target content was picked based on three methods of Twitter API data collection :

- **Streaming API** : that allows getting the tweets in real time. Selecting tweets in Arabic meeting Morocco as location requirement was performed. We managed to extract 8975 valid tweets from 23743 tweets collected.
- **User timeline API** : that returns the most recent tweets posted by a specific user. In this stage, we used SocialBakers[3] to determine the most

[2] https://developer.twitter.com/en/docs/twitter-api
[3] https://www.socialbakers.com

active Moroccan Twitter accounts. We managed to extract 5371 valid tweets from 10485 collected tweets from a list of 29 user timeline.

- **Search API**: that retrieves a tweets that match a particular query. At this stage, we resorted to the Trendsmap[4] to extract the hottest hashtags in Morocco. We have around 950 separate hashtag which are again used to collect tweets. We managed to extract 3654 valid tweets after filtering the Arabic tweets.

The final corpus consists of 18000 valid tweets based on 36114 tweets collected. 9750 were published by men and 7550 by women. Of 700 tweets could not be identified the gender of the author due to lack of information. Table 3 shows the distribution of tweets collected.

Table 3: Statistics on valid tweets annotated.

MSA	DA	Total
9640	8360	18000

3.2 Data annotation

The collected tweets were tagged by two native speakers (ourselves). A first step of double labeling involving 2500 tweets was carried out. It allowed us to familiarize ourselves with the task of labeling. The rest of the tweets were tagged individually. We calculated the inter-annotator agreements on the 2500 duplicate tagged messages (IAA = 0.97). The tweets have been tagged according to the polarity (Positive, Negative, Neutral and Mixed) and language of the tweets (Standard Arabic, Dialectal Arabic). The labeling was done through a web application that we have developed to accomplish this task.

3.3 Data properties

The corpus has been categorized into four classes of sentiment, namely Positive, Negative, Neutral and Mixed (positive + negative), and two classes of languages used MSA and DA. The corpus characteristics are shown in Figure 2 and Table 4.

3.4 Lexicon construction

The construction of emotions lexicon is a very difficult task that conditions the success of the lexicon-based approach. The challenge arises from the complexity of the Arabic language and the large number of words to be taken into account.

[4] https://www.trendsmap.com

Table 4: Details on valid tweets collected.

Number of tweets	18000
Number of retweets	4076
Number of distinct users	2354
Data collection period	01-Jan — 27-Apr 2021

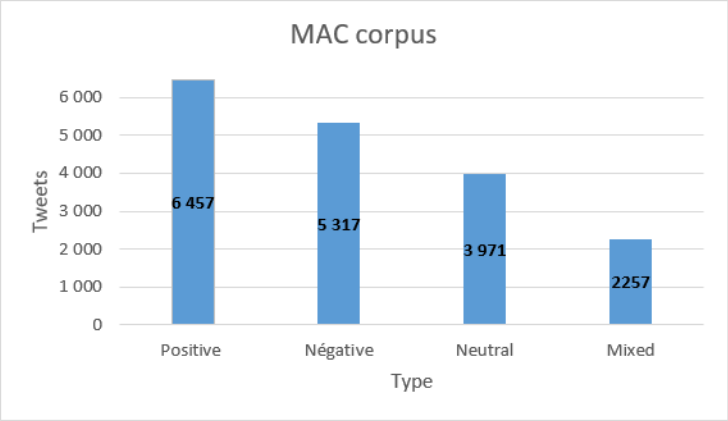

Fig. 2: Classes distribution.

In addition, determining the polarity of many words can be very difficult for many reasons, such as the different meanings and connotations of each word depending on the context and cultural background of the person who publish the tweet. The table below shows an example where a word can have multiple meanings / polarities depending on the context :

Table 5: Statistics on valid tweets annotated.

Tweet	Word	Polarity
Ar: أنت فى تقدم **Eng:** You are in progress	تَقَدّم	Positive
Ar: تقدم إلى الأمام **Eng:** Forward	تَقَدّم	Neutral

The constructed lexicon-dictionary in this study is created automatically from the annotated corpora. It consists of about 30.000 Moroccan Arabic term, where each word is assigned a polarity (positive, negative or neutral). Table 6 provides statistics about the constructed dictionary.

Table 6: Statistics about the constructed dictionary.

Positive	Negative	Neutral	Total
10671	2057	17272	30000

4 Experiments & Results

The SVM, Logistic Regression, CNN and LSTM classifier form the algorithms that we have chosen to conduct this benchmark study. In order to train the classifiers, a 5-fold stratified cross validation strategy was used.

We have proposed *two tasks*, the first is to *identify the used language* (MSA or DA) and the second centered on the *analysis of feelings*, given a tweet written in MSA or DA, this task consists of classifying it according to the feeling / emotion expressed by its author (positive, negative, neutral or mixed).

4.1 Task 1 : Language identification

Tables 7 shows the different scenarios settings that we have evaluated. The first column represents the used classifier name. The second column indicates the used features and whether a Stopwords filter was used (1) or removed (0). The third column shows the obtained accuracy. The best result by using the word embeddings representation is of 91.27%, obtained by using the LSTM classifier keeping the stop words, and 89.78% by removing them is acheived by the CNN.

Table 7: Results of identification of the used language.

Model	Features	Stop words	Accuracy
LSTM	Word embeddings	0	**91.27**
		1	89.23
CNN	Word embeddings	0	89.78
		1	89.16
SVM	TF-IDF	0	89.13
		1	86.30
LR	TF-IDF	0	88.64
		1	87.08

4.2 Task 2 : Sentiment analysis

Table 8 shows the obtained results for the second task. The best result in terms of accuracy on the MSA corpus is 92.09% was obtained using the LSTM classifier. On the Dialectal corpus, the CNN reached an accuraccy score of 85.42%

using the lexicon-based approach, while LSTM attained 93.24% on the whole corpus(AS+DM) using the supervised approach (corpus-based).

Table 8: Evaluation results of the second task.

Model	Approach	Performance		
		AS	DM	AS+DM
CNN	Corpus	91.78	84.17	90.87
	Lexicon	90.85	**85.42**	89.25
LSTM	Corpus	**92.09**	83.36	**93.24**
	Lexicon	90.88	84.53	89.62
SVM	Corpus	84.75	67.80	88.05
	Lexicon	82.04	74.14	78.11
LR	Corpus	82.23	65.78	79.88
	Lexicon	81.08	71.77	77.96

5 Conclusion

This study set out to fill the gaps in the literature by constituting the largest and freely available[5] benchmark corpus of Moroccan tweets created for ASA. This paper described in detail the creation and pre-processing of the corpus and lexicon, explained the annotation details that were adopted in creating the MAC, and described the features of the corpus, which consists of 18000 Moroccan tweets and a lexicon dictionnary of 30000 distinct words. A serie of experiments was applied on the MAC to offer benchmark results for forthcoming works. Furthermore, this dataset could be expanded to include more valuable texts coming from Facebook and/or YouTube.

References

[1] A. Abdelli et al. "Sentiment Analysis of Arabic Algerian Dialect Using a Supervised Method". In: *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*. 2019, pp. 1–6. DOI: 10.1109/ISACS48493.2019.9068897.

[2] M. Garouani, H. Chrita, and J. Kharroubi. "Sentiment Analysis of Moroccan Tweets Using Text Mining". In: *Digital Technologies and Applications*. 2021, pp. 597–608. DOI: 10.1007/978-3-030-73882-2_54.

[3] A. el Abdouli, L. Hassouni, and H. Anoun. "Sentiment Analysis of Moroccan Tweets Using Naive Bayes Algorithm". In: *International Journal of Computer Science and Information Security,* 15 (2017).

[5] https://github.com/LeMGarouani/MAC

[4] S. Mihi et al. "MSTD: Moroccan Sentiment Twitter Dataset". In: *International Journal of Advanced Computer Science and Applications (IJACSA)* 11.10 (). DOI: 10.14569/IJACSA.2020.0111045.

[5] I. A. El-khair. "1.5 Billion Words Arabic Corpus". In: *arXiv:1611.04033 [cs]* (2016). arXiv: 1611.04033 [cs].

[6] N. A. Abdulla et al. "Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis". In: *International Journal of Information Technology* (2014), pp. 55–71. DOI: 10.4018/ijitwe.2014070104.

[7] M. A. Ayyoub, S. B. Essa, and I. Alsmadi. "Lexicon-Based Sentiment Analysis of Arabic Tweets". In: *International Journal of Social Network Mining* 2.2 (2015), p. 101. DOI: 10.1504/IJSNM.2015.072280.

[8] I. Guellil et al. "SentiALG: Automated Corpus Annotation for Algerian Sentiment Analysis". In: *Advances in Brain Inspired Cognitive Systems*. 2018, pp. 557–567. DOI: 10.1007/978-3-030-00563-4_54.

[9] M. Maghfour and A. Elouardighi. "Standard and Dialectal Arabic Text Classification for Sentiment Analysis". In: *Model and Data Engineering*. 2018, pp. 282–291. DOI: 10.1007/978-3-030-00856-7_18.

[10] A. Oussous et al. "ASA: A Framework for Arabic Sentiment Analysis". In: *Journal of Information Science* 46.4 (2020), pp. 544–559. DOI: 10.1177/0165551519849516.

[11] H. Rahab, A. Zitouni, and M. Djoudi. "SANA : Sentiment Analysis on Newspapers Comments in Algeria". In: *JKSU-Computer and Information Sciences* (2019). DOI: 10.1016/j.jksuci.2019.04.012.

[12] S. Mdhaffar et al. *Sentiment Analysis of Tunisian Dialects: Linguistic Ressources and Experiments*. 2017, p. 61. DOI: 10.18653/v1/W17-1307.

[13] H. Abdellaoui and M. Zrigui. "Using Tweets and Emojis to Build TEAD: An Arabic Dataset for Sentiment Analysis". In: *Computación y Sistemas* 22.3 (2018). DOI: 10.13053/cys-22-3-3031.

[14] R. Alfared and H. A. Alhammi. "A Topic-Based Twitter Sentiment Analysis Training Dataset for Libyan Dialect". In: *International Journal of Advanced Computer Science and Applications* (2019), pp. 4–6.

[15] 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*, pp. 2515–2519. DOI: 10.18653/v1/D15-1299.

[16] N. Al-Twairesh et al. "AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets". In: *Procedia Computer Science* 117 (2017), pp. 63–72. DOI: 10.1016/j.procs.2017.10.094.

[17] A. Al-thubaity et al. *A Saudi Dialect Twitter Corpus for Sentiment and Emotion Analysis*. 2018, p. 6. DOI: 10.1109/NCG.2018.8592998.