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

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

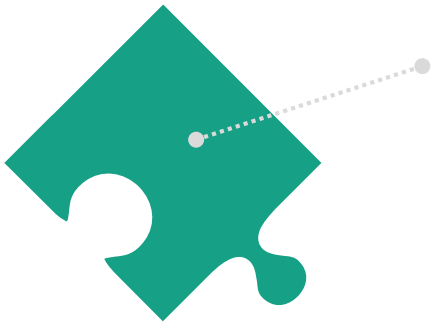
[1] *LISIC Laboratory, Univ. Littoral Cote d'Opale Calais, France*
[2] *LSIA Laboratory, Faculty of sciences and techniques Fez, USMBA, Morocco*

Motivation

With the advent of the web 2.0 and the explosion of data sources such as review platforms, blogs and microblogs, there has been a need to analyze millions of posts, tweets or reviews in order to find out what internet users think.

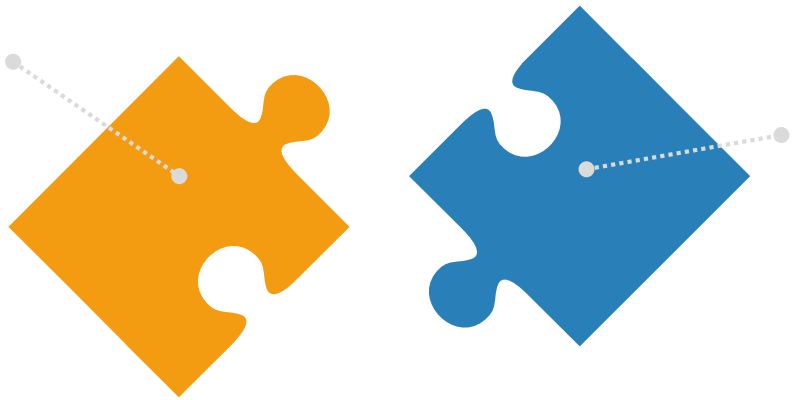
Motivation

The number of active social media users in Morocco has increased by **4M**[1] users over the past year, reaching the number of **22 million** social media users.



1- The research carried out on the analysis of the sentiment of tweets in Arabic is very limited, in particular Moroccan Arabic compared to other languages.

3- Morocco is thus ranked 9th among Arab countries with the highest number of users. .



2- The total lack of additional resources for Moroccan Arabic.

[1] https://www.statista.com/statistics/1172771/number-of-social-media-users-morocco/

PLAN

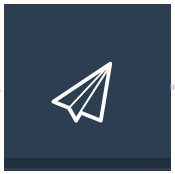
Introduction

Sentiment analysis

Background and literature review

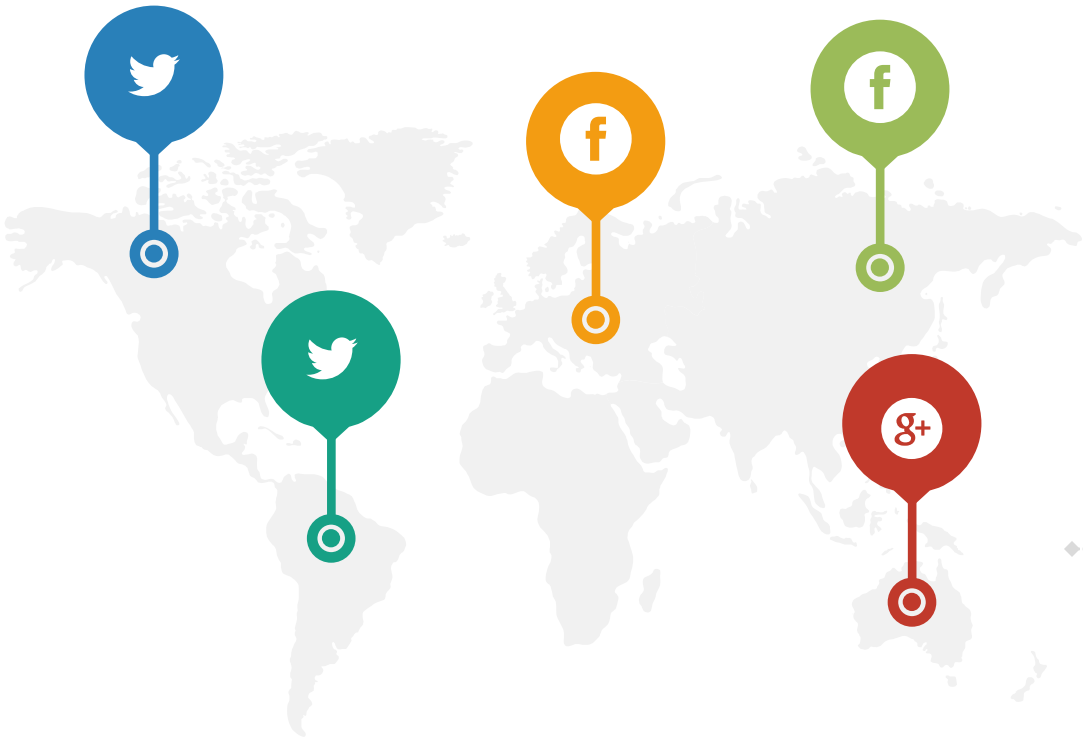
Benchmark corpus

Experiments and Results



Conclusion & Perspectives

Introduction



Social media

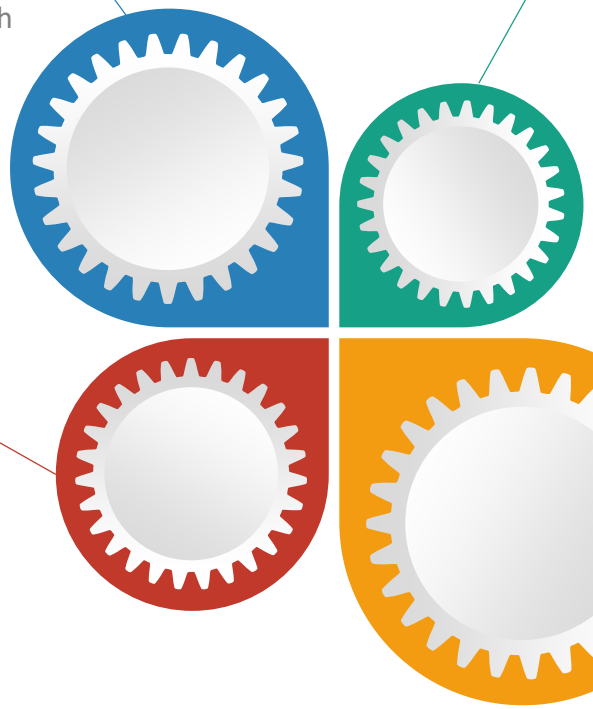
Facebook, Twitter, Instagram, LinkedIn, these social platforms are now part of everyday life. The data aspect of these social media, such as Twitter messages, generates a rich wealth of data about who is involved in communication.

This data plays an important role in decision making for many people and organizations.

Sentiment Analysis

Sentiment analysis

Refers to technologies for the automatic analysis of speech, written or spoken, in order to extract subjective informations such as judgments, evaluations or emotions.



Data Sources

- Review sites
- Blogs
- Micro-blogs: Twitter, Facebook…

Application areas

- Politics / political science
- Commercial
- Sociology
- Finance

Approaches

- Machine Learning Approach
- Lexicon-based / dictionary rule-based methods (Semantic orientation)

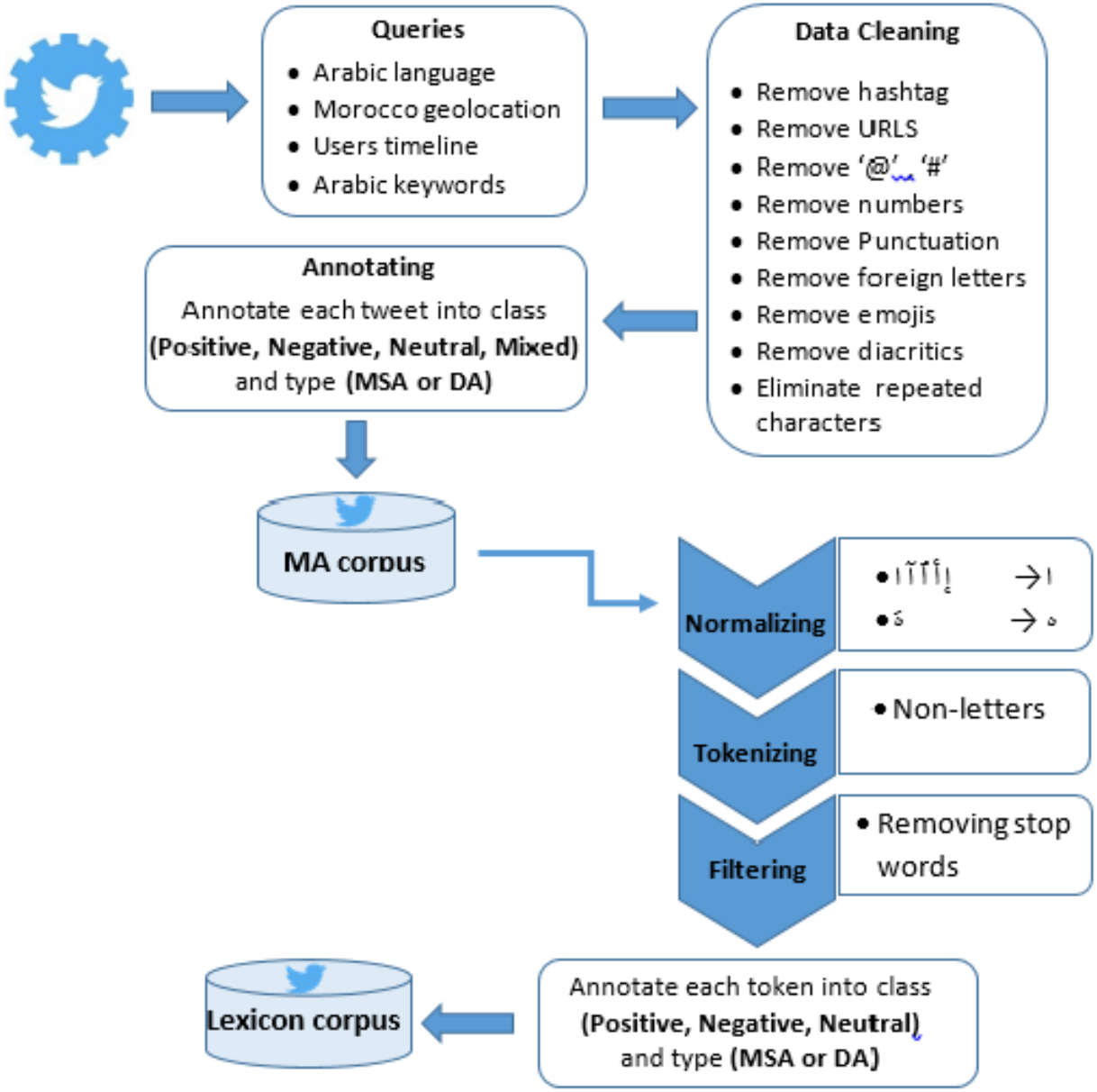
State of art

Maghrebian corpora

Dataset	Size	Arabic	Classes	Source	Year	Publicly Available
[15]	10006	Egyptian	4	Twitter	2015	✓
[13]	6m	Tunisian	2	Twitter	2017	✓
[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	✗

Table 1: Arabic corpora for sentiment analysis.

Benchmark corpus



I- Data collection

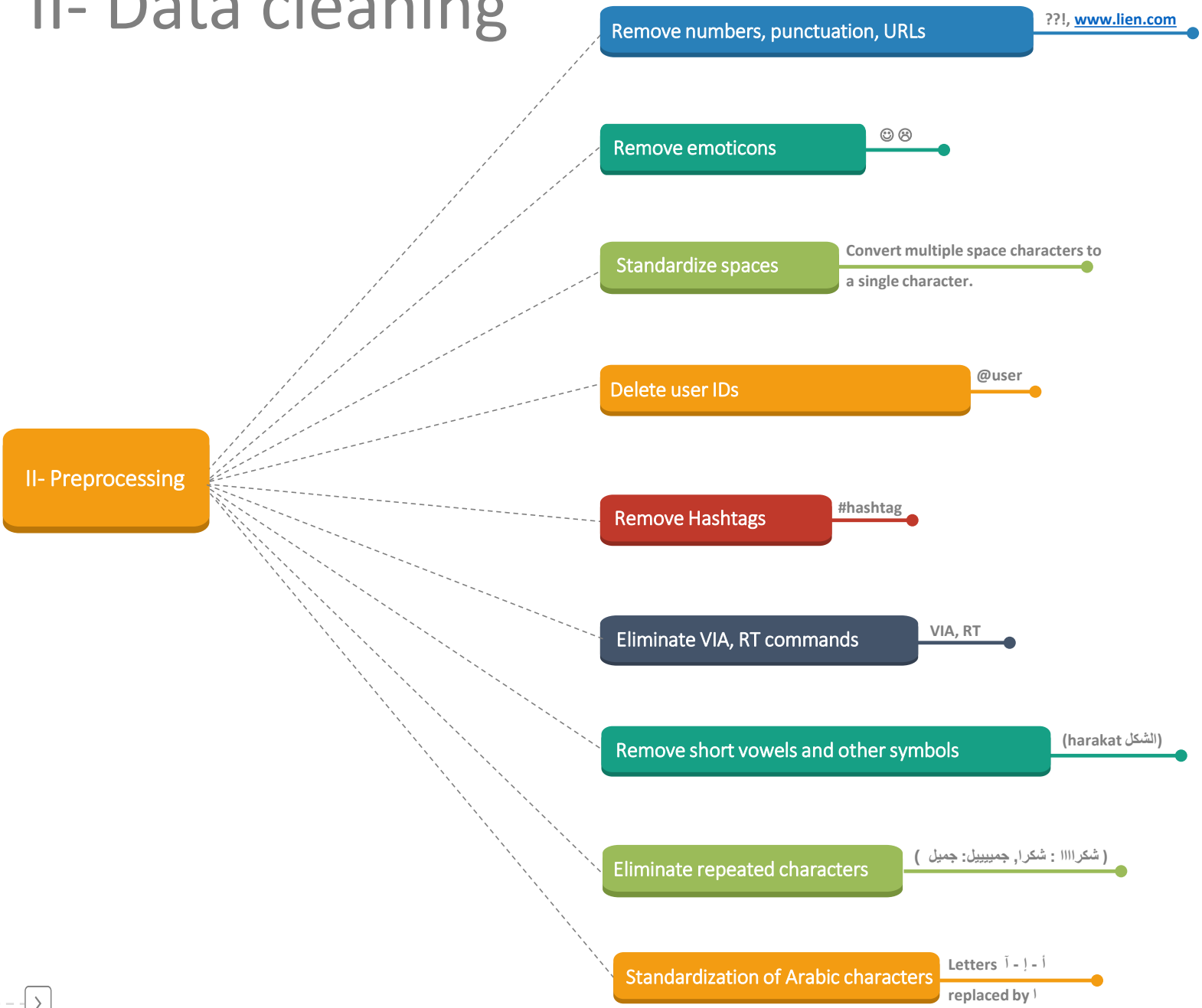
Final corpus

- The corpus consists of the total of 18000 valid tweets based on 36,114 tweets collected

Number of tweets collected	18000
Number of valid tweets	8360
Number of retweets	9640

Table 2: Statistics on the final corpus.

II- Data cleaning



III- Annotation

- The corpus was labeled by ourselves, our task is to determine the polarity (Positive, Negative, Neutral, Mixed) and the language of the tweets (SA or MD).
- The annotation was done through a web application

Tweet	Type	Class
Ar : توقع الخير و افتح صباحك بالتفاؤل و الأمل صباح النور En: Expect the good things and start your day with optimism and hope	Positive	SA
Ar : من المؤسف ان هذا حالنا الذي نعيشه الآن En: Unfortunately, this is our current situation	Negative	SA
Ar : تابعيني باش نقدر ندخلك En: Subscribe so that I can add you	Neutral	MD
Ar: رغم الصعوبات لي قاتلاني والمشاكل لي كنمر منها كنحاول نضحك ونقول الحمد لله En: Despite the difficulties and problems I have I try to laugh and thank God	Mixed	MD

Table 3: Example of annotated tweets

III- Annotation

The distribution of data according to their class and sentiment is shown in the following table:

SA	MD	Total
9 640	8360	18000

Table 4: Statistics on the corpus.

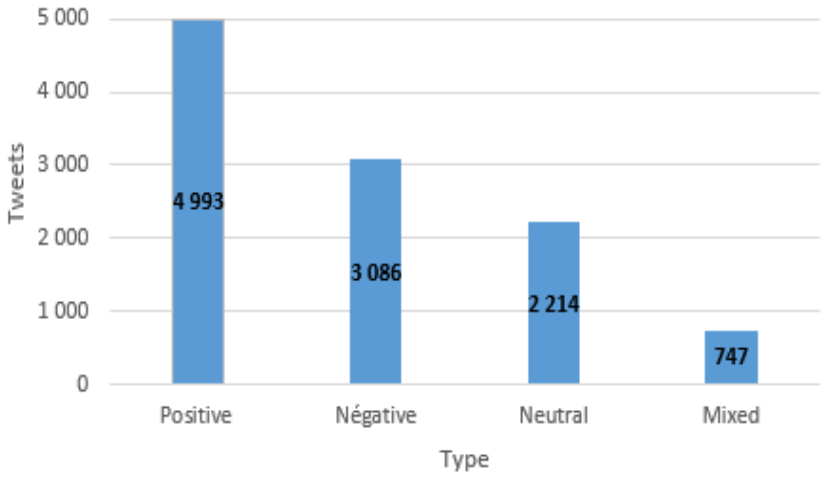


Figure 1: Distribution of feelings expressed in the SA corpus.

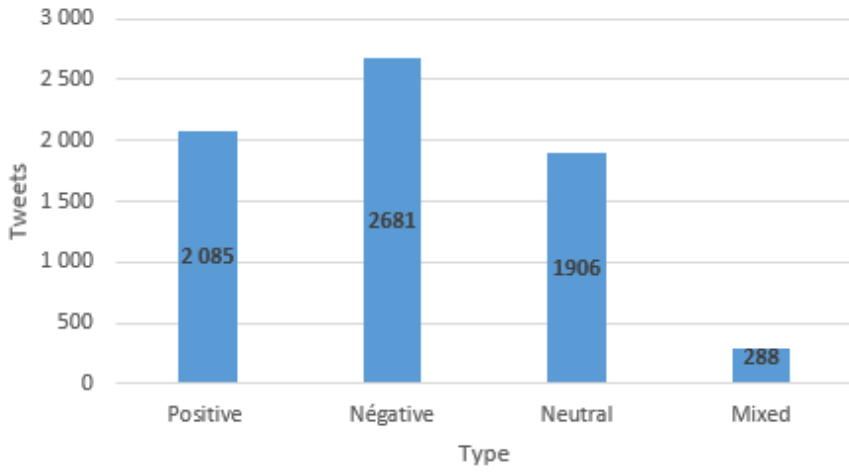


Figure 2: Distribution of feelings expressed in the MD corpus.

Lexicon construction

- Statistics on the built dictionary:

Positif	Négatif	Neutre	Total
2 630	2 057	13 995	18 683

Table 5: Lexicon extracted from the SA database .

Positif	Négatif	Neutre	Total
1 291	702	8 902	10 895

Table 6: Lexicon extracted from the MD database .

V- Classification

Classifiers used

1. Convolutional Neural Networks (CNN)
2. Short-term long-term memory networks (LSTMs)
3. Support Vector Machine (SVM)
4. Logistic regression (LR)

Analysis evaluation

Task 1 : Language identification

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
Logistic Regression	TF-IDF	0	88.64
		1	87.08

Table 6: Results of identification of the used language.

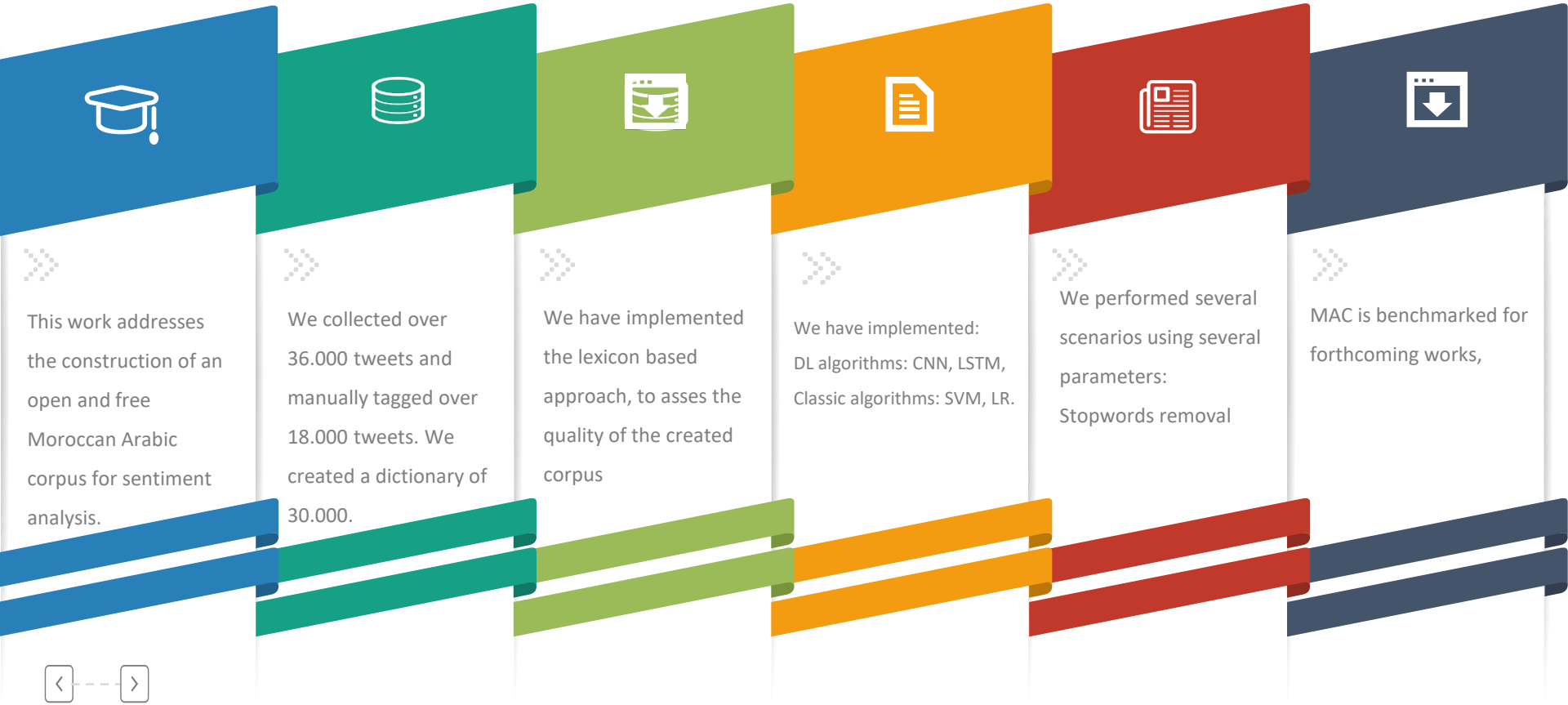
Analysis evaluation

Task 2 : Sentiment analysis

Model	Approach	Accuracy		
		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
Logistic Regression	Corpus	82.23	65.78	79.88
	Lexicon	81.08	71.77	77.96

Table 7: Evaluation results of the second task.

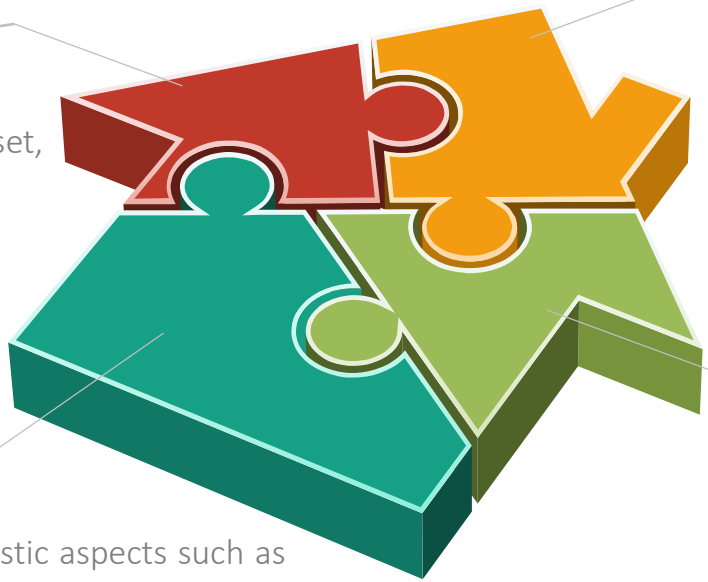
Conclusion


This work addresses the construction of an open and free Moroccan Arabic corpus for sentiment analysis.
We collected over 36.000 tweets and manually tagged over 18.000 tweets. We created a dictionary of 30.000.
We have implemented the lexicon based approach, to asses the quality of the created corpus
We have implemented:
DL algorithms: CNN, LSTM,
Classic algorithms: SVM, LR.
We performed several scenarios using several parameters:
Stopwords removal
MAC is benchmarked for forthcoming works,

Perspectives

The next planned steps include:

1. Increase in the size of the dataset, in particular the DM corpus



2. Discussion of the issue of imbalance between data sets and text.

3. Add more parameters more features and classifiers.

4- The involvement of other linguistic aspects such as the type of words (subject, verb, adjectives, etc.) which can improve the process of sentiment analysis.

THANK YOU FOR YOUR ATTENTION

To your questions

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

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

[1] *LISIC Laboratory, Univ. Littoral Cote d'Opale Calais, France*
[2] *LSIA Laboratory, Faculty of sciences and techniques Fez, USMBA, Morocco*