Sentiment analysis of Moroccan tweets using text mining

Intelligent Systems and Applications Laboratory (LSIA)

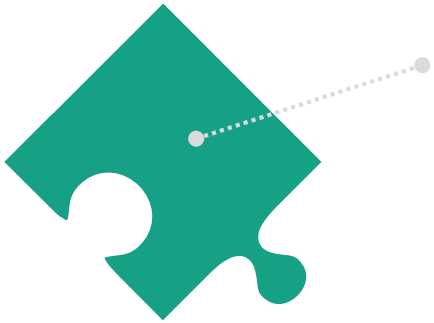
- @ Garouani Moncef
- @ Chrta Hanae
- @ Kharrouabi Jamal

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.

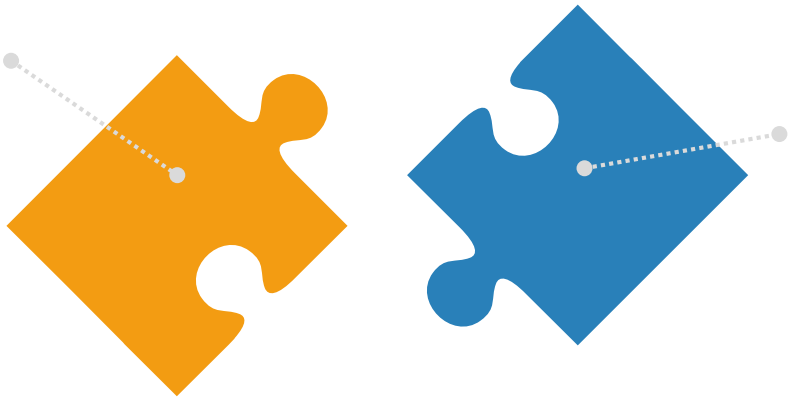
Motivation

The number of Moroccan active users of the Twitter platform has increased by **500.000** users over the past year, reaching the number of + **2 000 000** 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.

PLAN

Introduction

Sentiment analysis

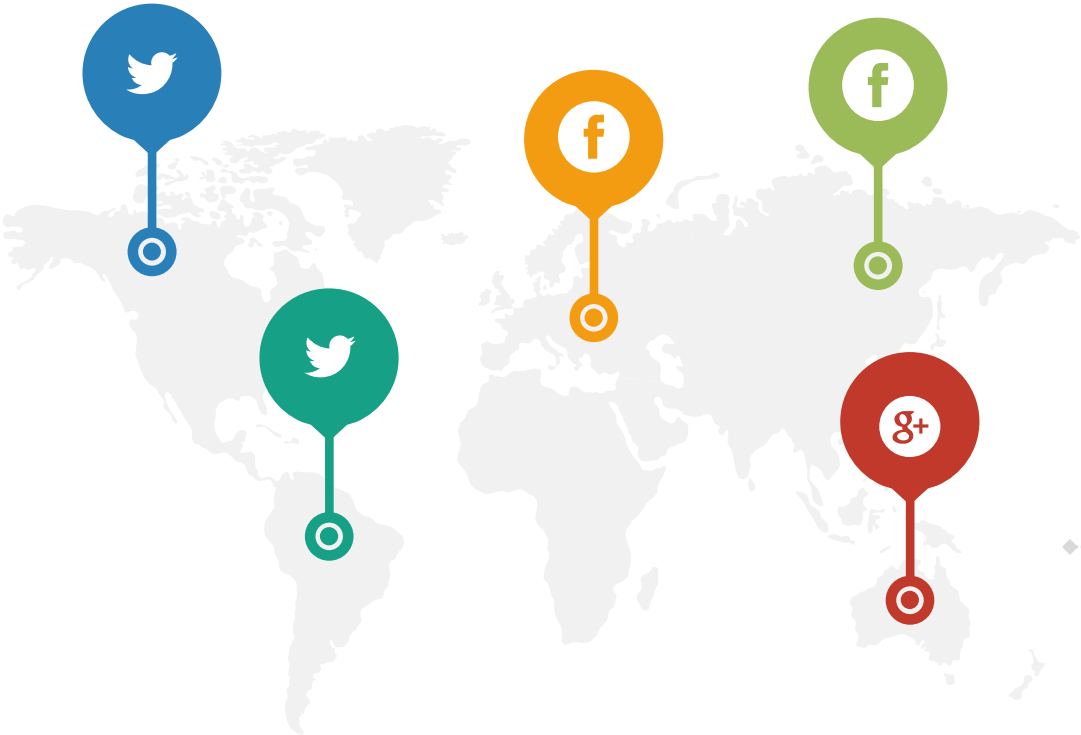
Background and literature review

Proposed system

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.

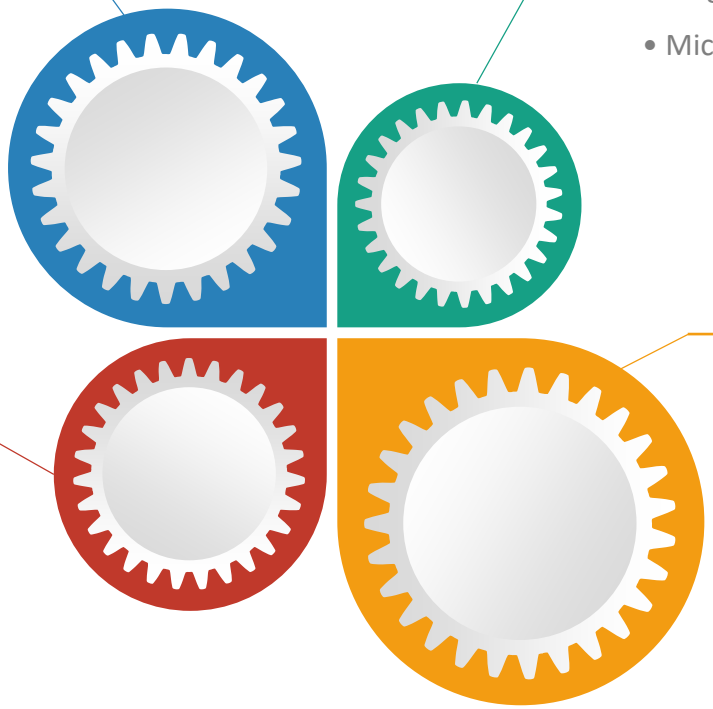
Sentiment Analysis

Sentiment analysis

Refers to technologies for the automatic analysis of speech, written or spoken, in order to extract subjective information 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

Sentiment analysis

Alomari, et al. 2017

Proposed a **Jordanian dialect** corpus of Eightine hundred **1,800 tweets** (900 positive and 900 negative) to compare the performance of SVMs and naive bayes in sentiment analysis using different preprocessing methods. The classification accuracy was **88.72%** using SVMs.

Shoukry et Rafea, 2012

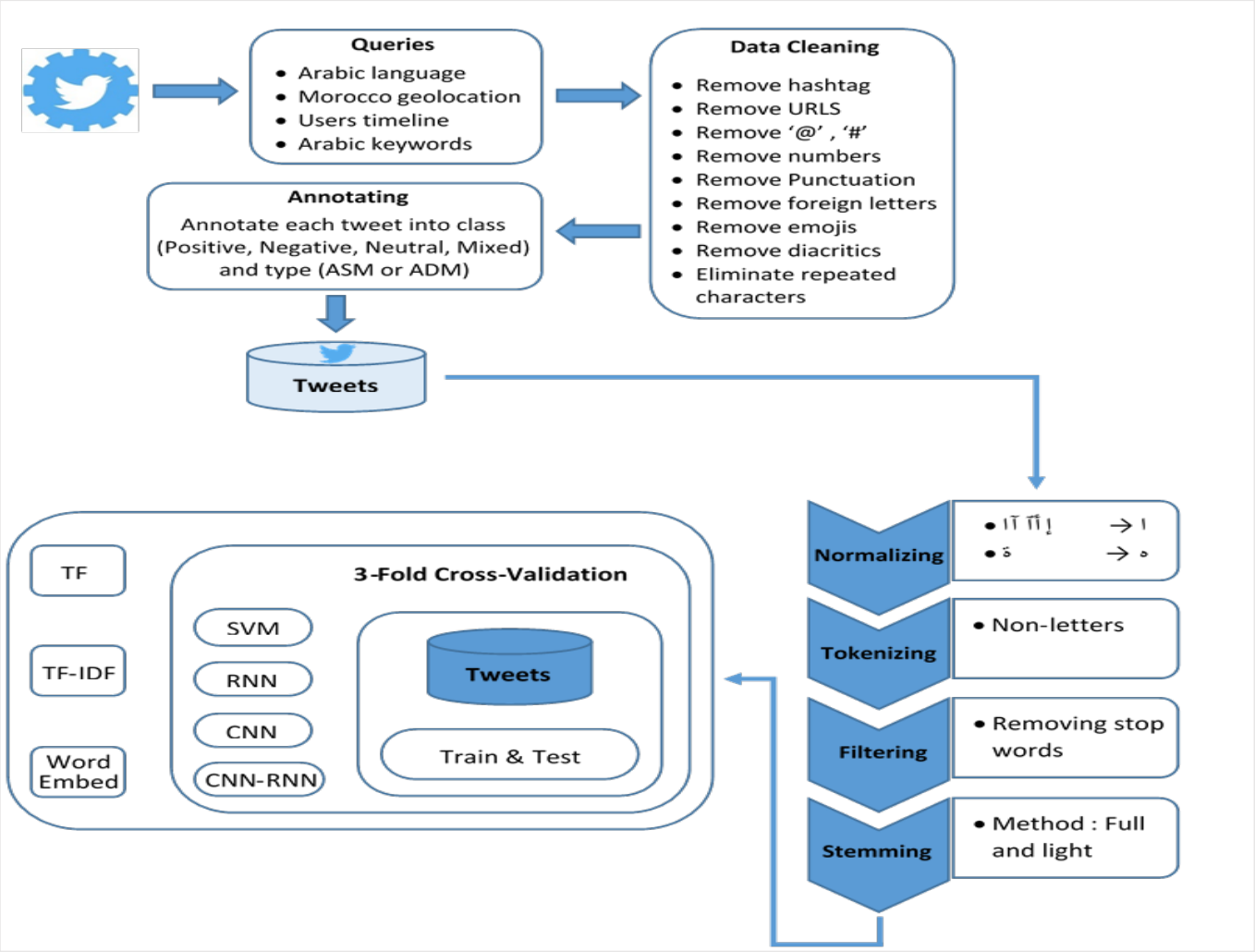
Collected **1000 tweets** on several hot topics. In their experiment, the SVM algorithms were applied using the software 'Weka Suite' for the classification process in the two approaches used (ML and SA). The classification accuracy by the **machine learning approach was 78.8%** and that of the **lexical approach 75.9%.** However, their data set was relatively small.

Abdeljalil EL ABDOULI et Al. 2017

Worked on a supervised classification (naive bayes) using Seven hundred **700 tweets** (positive and negative) based on emojis to analyze feelings. Classification accuracy was **69%** using **SVMs,**

Framework

We can summarize our approach to classification of feelings by the following diagram:



I- Data extraction

Final corpus

- The corpus consists of the total of 13,550 valid tweets based on 36,114 tweets collected

thirteen thousand five hundred fifty thirty-six thousand

Number of tweets collected	**36 114**
Number of valid tweets	13 550
Number of distinct users	3 602

Table 3: Statistics on the final corpus.

II- Preprocessing

- Remove numbers, punctuation, URLs — ??!, www.lien.com
- Remove emoticons — ☺☹
- Standardize spaces — Convert multiple space characters to a single character.
- Delete user IDs — @user
- Remove Hashtags — #hashtag
- Eliminate VIA, RT commands — VIA, RT
- Remove short vowels and other symbols — (harakat الشكل)
- Eliminate repeated characters — (شكراااا : شكرا, جمييييل: جميل)
- Standardization of Arabic characters — Letters أ - إ - آ replaced by ا

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 (AS or DM).
- 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	AS
Ar : من المؤسف ان هذا حالنا الذي نعيشه الآن En: Unfortunately, this is our current situation	Negative	AS
Ar : تابعيني باش نقدر ندخلك En: Subscribe so that I can add you	Neutral	DM
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	DM

Table 4: Example of annotated tweets

III- Annotation

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

AS	DM	Total
9 640	3 807	13 550

Table 5: Statistics on the corpus.

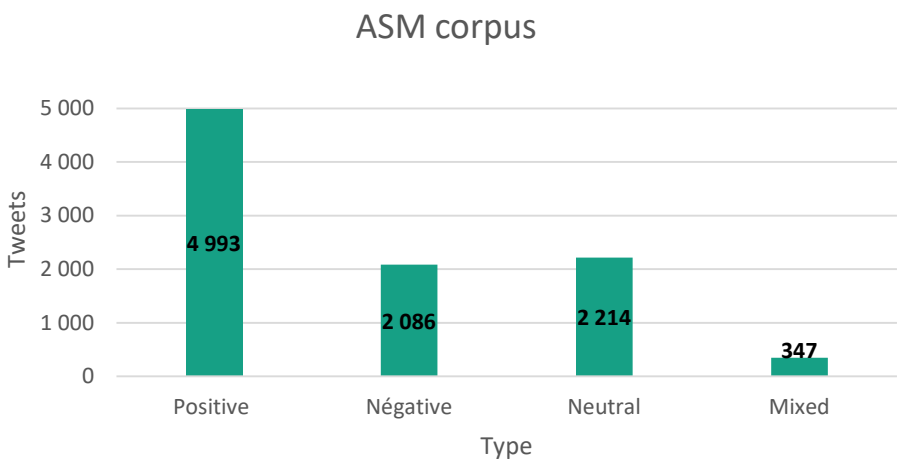


Figure 2: Distribution of feelings expressed in the AS corpus.

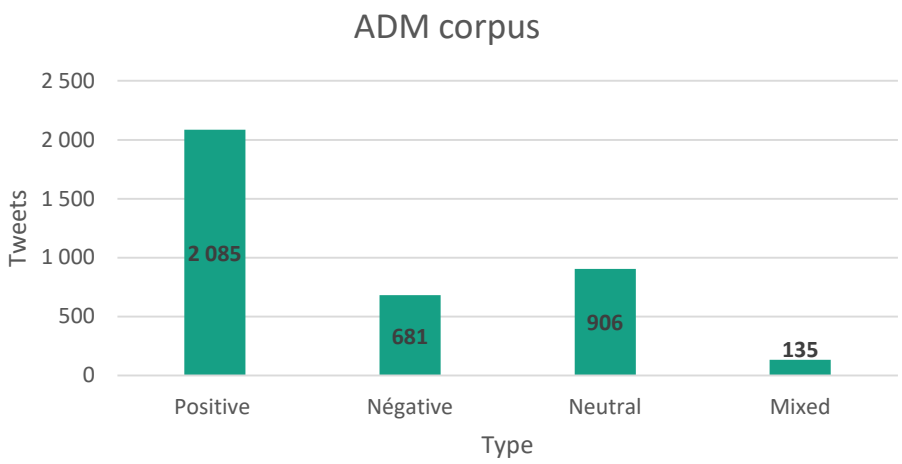
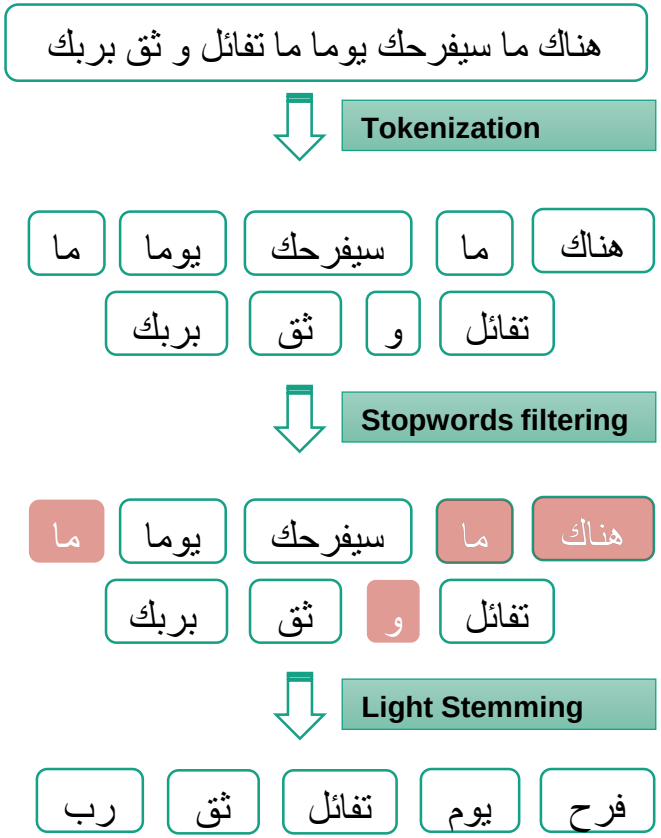


Figure 3: Distribution of feelings expressed in the DM corpus.

IV- Text preprocessing and transformation process.



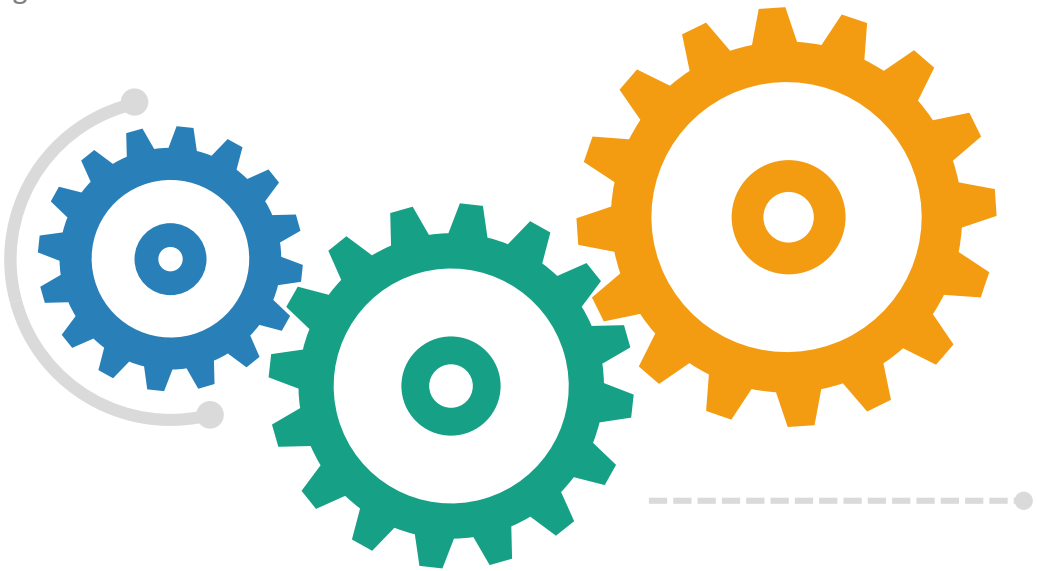
V- Classification

Classifiers used

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

Analysis evaluation

we have proposed 2 tweet analysis tasks focused on the analysis of feelings and the language used:



Task 1: Identification of the language.

Given a tweet, this task is to identify the language used (AS or DM).

Task 2: Classification of tweets according to their sentiment polarity.

Given a tweet written in AS or DM, this task consists of classifying it according to the feeling / emotion expressed by its author, in: positive, negative, neutral or mixed.

Analysis evaluation

Approche Machine Learning

I.1 Results of the first task

Modèles	Features		Accuracy
LSTM	Word Embedding	Avec sw	**88.86**
		Sans sw	86.95
CNN-LSTM	Word Embedding	Avec sw	88.43
		Sans sw	87.06
CNN	Word Embedding	Avec sw	88.32
		Sans sw	**87.37**
SVM	TF-IDF	Avec sw	88.39
		Sans sw	87.25
LR	TF-IDF	Avec sw	88.17
		Sans sw	87.02

Table 7: Results of the classification of the type of language used.

Analysis evaluation

I.2 Results of the second task

Modèles Deep Learning	Features		Accuracy			Accuracy all	
			AS	DM	AS_DM	AS	DM
CNN	Word Embedding	Avec SW	**91.78**	84.17	89.56	91.62	**85.37**
		Sans SW	91.39	83.82	89.08	91.12	84.78
LSTM		Avec SW	**92.09**	83.36	89.60	91.80	84.69
		Sans SW	91.64	82.87	89.49	91.36	**85.04**
CNN-LSTM		Avec SW	**91.83**	81.75	89.46	91.50	**85.55**
		Sans SW	91.67	82.00	88.93	91.46	85.26

Table 8: Results of Deep Learning + Word Embedding classifiers on Moroccan tweets.

Analysis evaluation

I.2 Results of the second task

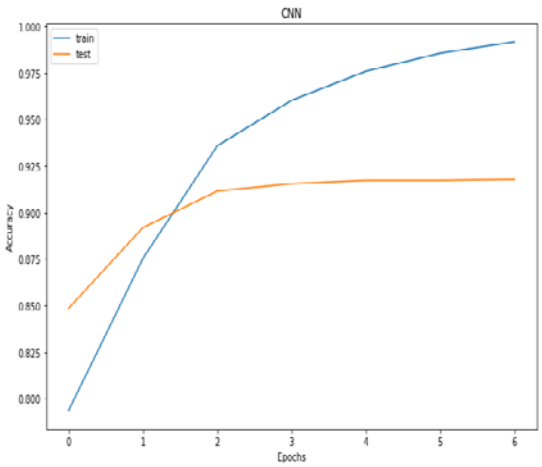
Modèles Classique	Features			Accuracy			Accuracy all	
				AS	DM	AS-DM	AS	DM
SVM	TF-IDF	Uni-gram	Avec SW	83.50	67.01	78.70	82.06	70.08
			Sans SW	82.30	66.75	78.77	82.78	68.07
		Bi-grams	Avec SW	**84.75**	67.80	**80.05**	83.84	70.00
			Sans SW	83.47	68.68	79.41	82.54	**71.40**
		Tri-grams	Avec SW	84.15	67.40	80.00	83.33	70.35
			Sans SW	83.33	67.54	79.86	83.94	69.38
Logistic Regression		Uni-gram	Avec SW	**82.23**	64.38	78.18	82.03	69.82
			Sans SW	82.07	65.78	**78.55**	82.44	68.59
		Bi-grams	Avec SW	81.27	62.36	77.52	80.93	68.77
			Sans SW	81.14	60.88	78.08	81.14	70.26
		Tri-grams	Avec SW	81.31	62.10	77.96	81.58	68.68
			Sans SW	80.73	61.22	77.54	81.00	68.72

Table 9: Results of classic classifiers + TF-IDF on Moroccan tweets.

Analysis evaluation

Approche Machine Learning

I.1 Results of the first task



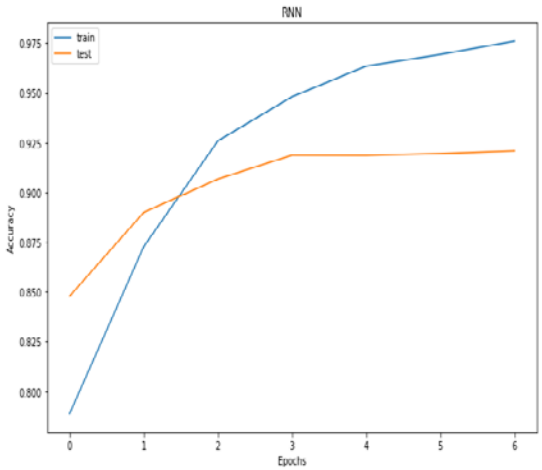


Figure 9: Comparison of the different classifiers on the AS corpus.

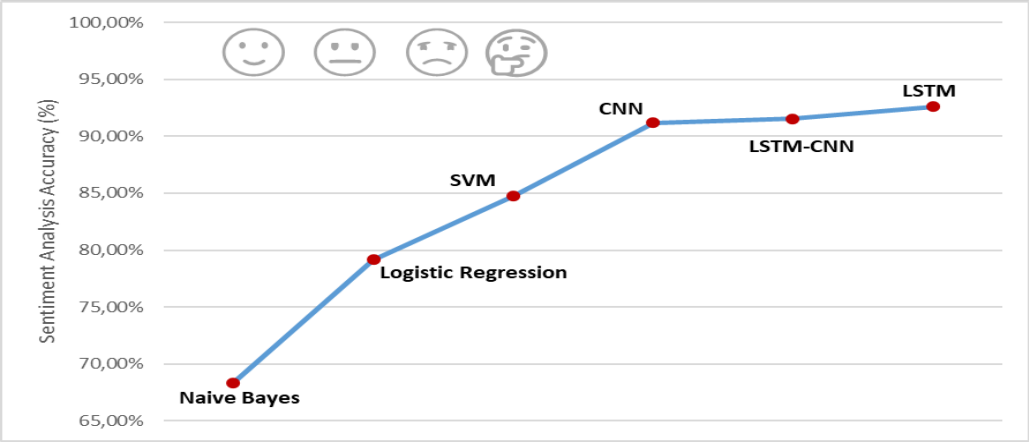


Figure 10: Average performance results of CNN and LSTM classifiers.

Conclusion

This work addresses sentiment analysis in Moroccan Arabic tweets.

We collected over 36,000 tweets and manually tagged over 13,550 tweets.

We have implemented the Machine Learning based approach for sentiment analysis

We have implemented:
DL algorithms: CNN, LSTM, CNN-LSTM
Classic algorithms: SVM, LR.

We performed several scenarios using several parameters:
N-grams, Stopwords removal, TF-IDF, and Word Embedding.

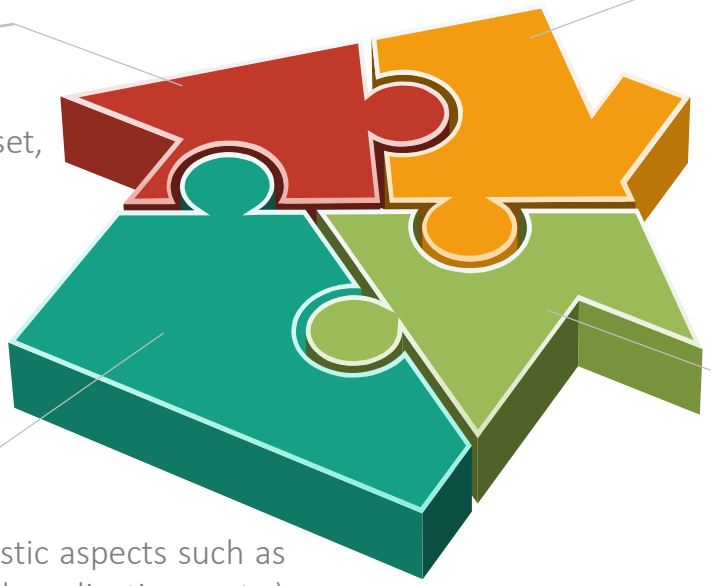
Our system achieves convincing results.
The system achieves an average precision of 96% for the two corpora.

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

Sentiment analysis of Moroccan tweets using text mining

Intelligent Systems and Applications Laboratory (LSIA)

- @ Garouani Moncef
- @ Chrta Hanae
- @ Kharrouabi Jamal