THE INTERNATIONAL CONFERENCE
ICI2C
ON INFORMATION, COMMUNICATION & CYBERSECURITY

# Towards a new Lexicon-Based features vector for Sentiment Analysis: Application to Moroccan Arabic tweets

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

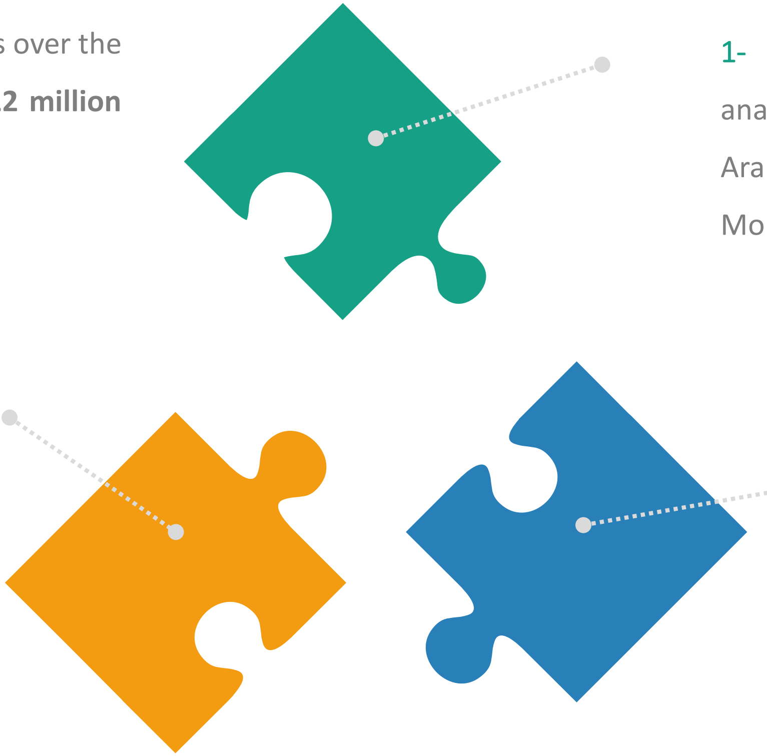
[1] *LISIC Laboratory, Univ. Littoral Cote d'Opale Calais, France*
[2] *LSIA Laboratory, Faculty of sciences and techniques Fez, USMBA, Morocco*

# 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.

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

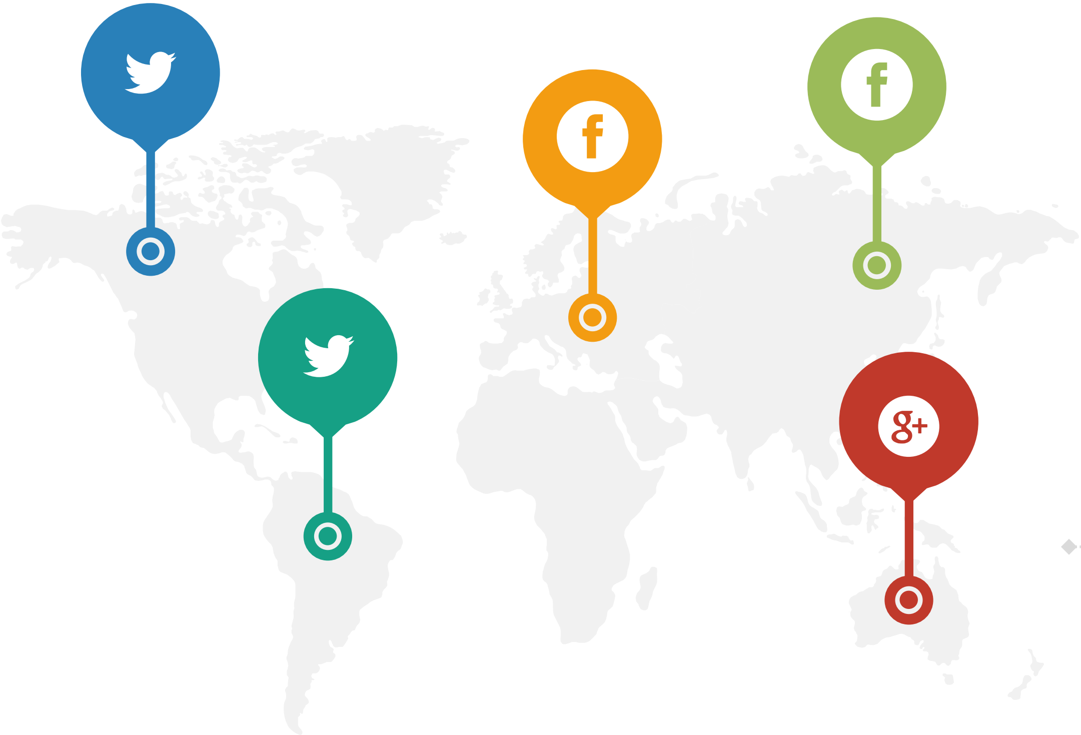
3- Morocco is thus ranked 9th among Arab countries with the highest number of users. .

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

# PLAN

- Introduction
- Sentiment analysis
- Background and literature review
- Proposed approach
- 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.

120K

530K

319K

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

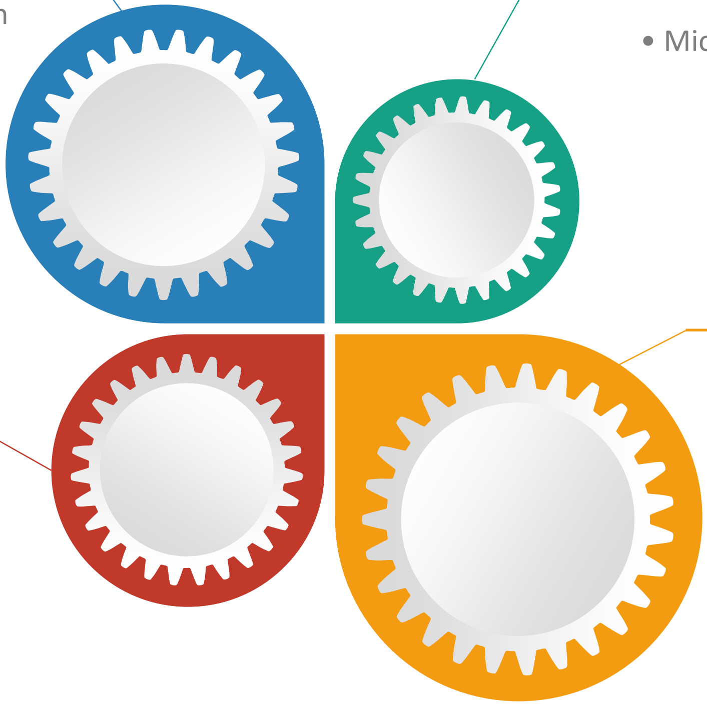
# Sentiment Analysis

## 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…



## Approaches

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

## Application areas

- Politics / political science
- Commercial
- Sociology
- Finance

# State of art

## Sentiment analysis

### Abdulla et al. 2014

Proposed a **domain-based lexicon approach** to deal with Arabic text (SA and colloquial Arabic). They created two lexicons for every domain (books, movies, society, politic, etc.), one for positive words and another for negative ones from a corpus of **1080** reviews compiled from different social networks. Their approach achieved an accuracy of **90%**.

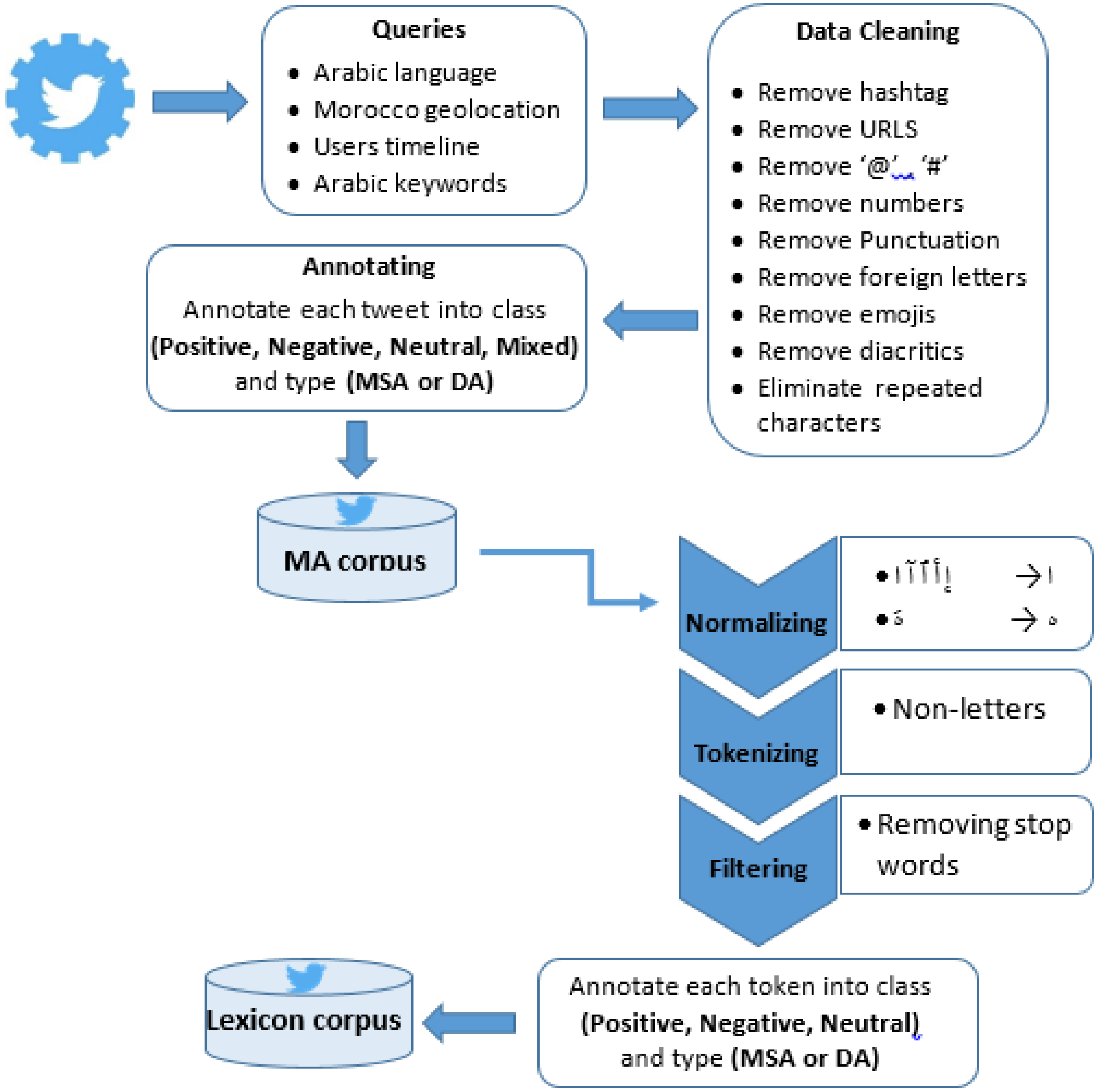
### Abdeljalil EL ABDOULI et Al. 2017

Discussed the sentiment analysis for Jordanian tweets, and built a tool for extracting the polarity of unstructured text where a weight representing the polarity is assigned to each word in the lexicon (+1 and -1 for positive and negative words, respectively).

### Al-Ayoub et al. 2015

Proposed an **unsupervised technique** for sentiment analysis of Arabic tweets. The first step of their technique was collecting tweets and applying preprocessing methods (i.e. stemming and stop-word removal). Next, a sentiment lexicon was constructed with polarity scores between **0** and **100**. Scores less than **40** indicated ***negative*** sentiment, between **40** and **60** corresponded to ***neutral***, while scores from **60** to **100** indicate ***positive*** sentiment. Finally, all these scores were combined to compute the sentiment score of the text. This technique has achieved **86.89%** in overall accuracy.

# Framework



**Queries**

- Arabic language
- Morocco geolocation
- Users timeline
- Arabic keywords

**Data Cleaning**

- Remove hashtag
- Remove URLS
- Remove '@', '#'
- Remove numbers
- Remove Punctuation
- Remove foreign letters
- Remove emojis
- Remove diacritics
- Eliminate repeated characters

**Annotating**

Annotate each tweet into class **(Positive, Negative, Neutral, Mixed)** and type **(MSA or DA)**

- إأٱآا → ا
- ة → ه

- Non-letters

- Removing stop words

Annotate each token into class **(Positive, Negative, Neutral)** and type **(MSA or DA)**

# I- Data collection

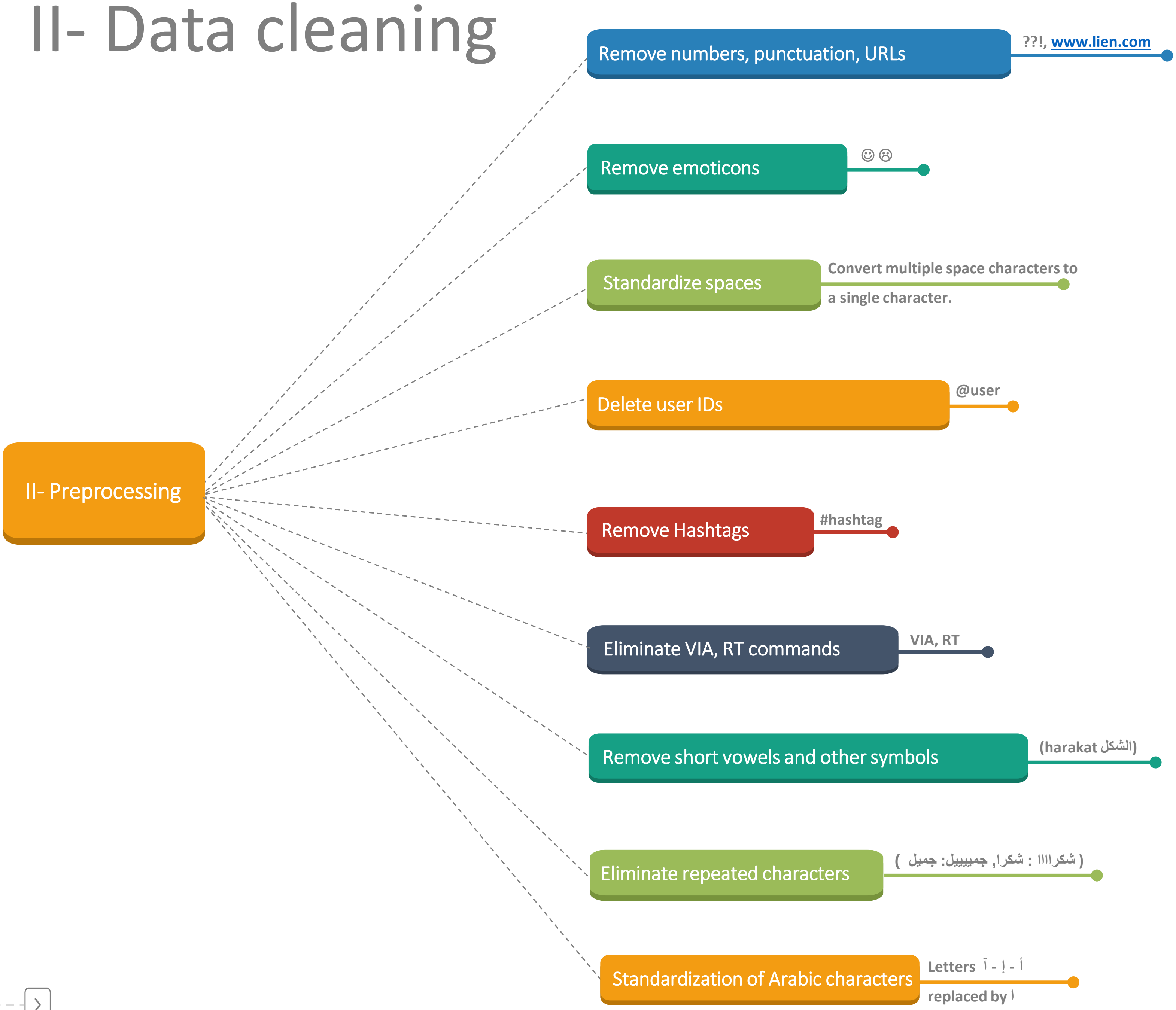
Final corpus

- The corpus consists of the total of 18,000 valid tweets based on 36,114 tweets collected

| | |
|---|---|
| **Number of tweets collected** | **36 114** |
| **Number of valid tweets** | 18.000 |
| **Number of distinct users** | 3 602 |

Table 1: Statistics on the collected corpus.

# II- Data cleaning



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<br>En: Expect the good things and start your day with optimism and hope | Positive | AS |
| Ar : من المؤسف ان هذا حالنا الذي نعيشه الآن<br>En: Unfortunately, this is our current situation | Negative | AS |
| Ar : تابعيني باش نقدر ندخلك<br>En: Subscribe so that I can add you | Neutral | DM |
| Ar :رغم الصعوبات لي قاتلاني والمشاكل لي كنمر منها كنحاول نضحك ونقول الحمد لله<br>En: Despite the difficulties and problems I have I try to laugh and thank God | Mixed | DM |

Table 2: 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** | 3 807 | 13 550 |

Table 3: Statistics on the corpus.

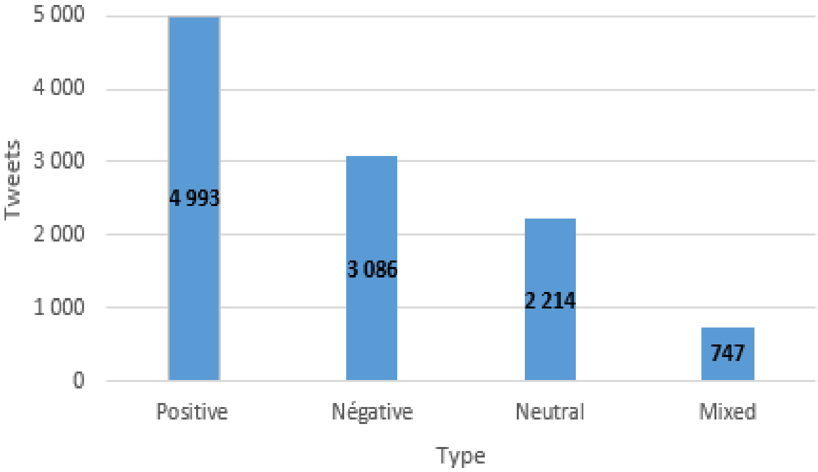


Figure 1: Distribution of feelings expressed in the AS corpus.

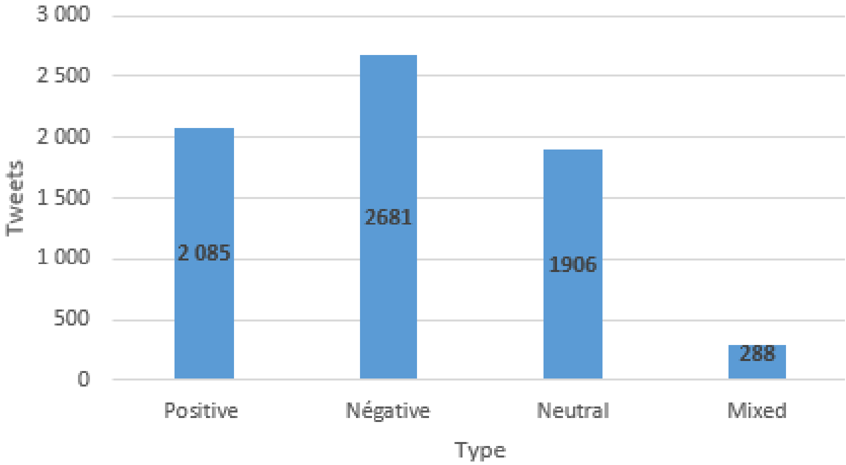
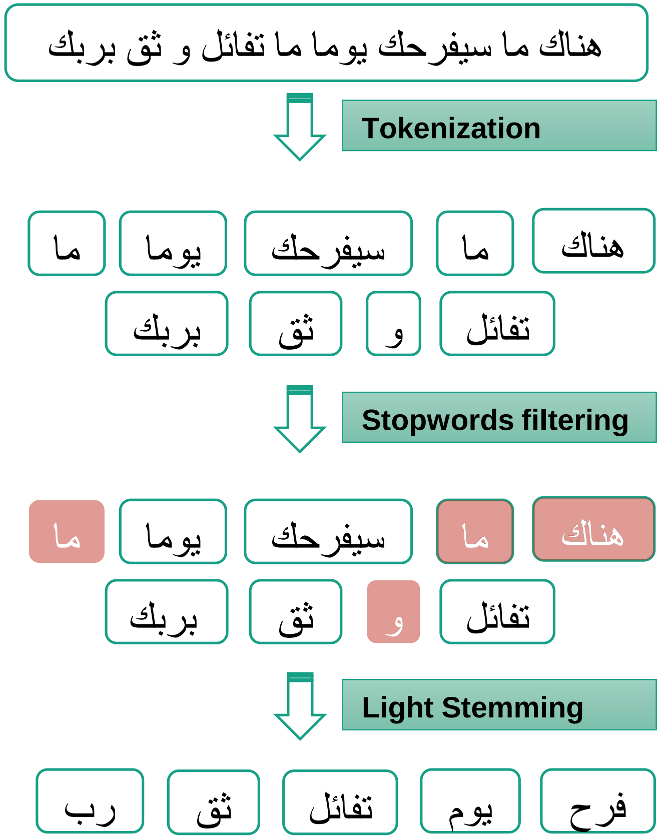


Figure 2: Distribution of feelings expressed in the DM corpus.

# IV- Text preprocessing and transformation process.



هناك ما سيفرحك يوما ما تفائل و ثق بربك

Tokenization

هناك ما سيفرحك يوما ما

تفائل و ثق بربك

Stopwords filtering

هناك ما سيفرحك يوما ما

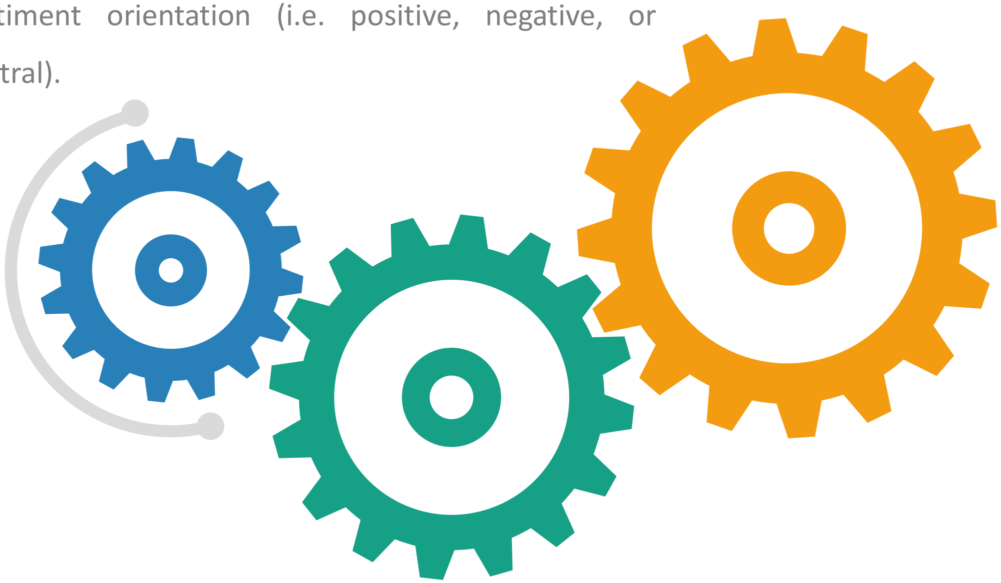
تفائل و ثق بربك

Light Stemming

فرح يوم تفائل ثق رب

# Lexicon-based approach

In sentiment analysis, *lexicons* are a synonym for *dictionaries*, except lexicons is sentiment analysis contain polarities along with the words instead of their definitions. That is, every word has an associated sentiment orientation (i.e. positive, negative, or neutral).



## Constructed lexicon

The adopted lexicon in this study is created automatically from the annotated corpus. It consists of about 30.000 Moroccan Arabic term, where each word is assigned a polarity (positive, negative or neutral).

# Lexicon construction

- Statistics on the built dictionary:

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

Table 4: Lexicon extracted from the SA database .

| Positif | Négatif | Neutre | Total |
|---|---|---|---|
| 1 291 | 702 | 8 902 | 10 895 |

Table 5: Lexicon extracted from the MD database .

# Lexicon-Based features vector

- To classify the tweet, a score is calculated for each sentiment (positive, negative and neutral) to build a vector that will represent the tweet, as follows:

$$Weight_{class} = \frac{Number\ of\ words\ of\ a\ class\ in\ the\ tweet}{Total\ number\ of\ words\ in\ the\ tweet}$$

- The final values of the weights determine the polarity of the whole tweet, representing it as a vector :

| | | |
|---|---|---|
| $score_{positif}$ | $score_{négatif}$ | $score_{neutre}$ |

Tweet Vector

| | | |
|---|---|---|
| $score_{positif}$ | 0 | $score_{neutre}$ |

Positive Tweet

| | | |
|---|---|---|
| 0 | $score_{négatif}$ | $score_{neutre}$ |

Negative Tweet

| | | |
|---|---|---|
| 0 | 0 | $score_{neutre}$ |

Neutral Tweet

| | | |
|---|---|---|
| $score_{positif}$ | $score_{négative}$ | $score_{neutre}$ |

Mixed Tweet

# 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

| Model | Stop words | Accuracy | | |
|---|---|---|---|---|
| | | AS | DM | AS_DM |
| CNN | with sw | 90.80 | **85.42** | **89.25** |
| | without sw | **90.85** | 85.30 | 89.14 |
| LSTM | with sw | **90.88** | 84.53 | 89.62 |
| | without sw | 90.63 | 84.02 | 88.66 |
| SVM | with sw | **82.04** | **74.14** | **78.11** |
| | without sw | 81.49 | 73.25 | 77.80 |
| Logistic Regression | With sw | **81.08** | 71.77 | **77.96** |
| | Without sw | 80.63 | 71.51 | 77.54 |

Table 6: Evaluation results of the proposed vector representation.

# Conclusion

This work addresses sentiment analysis in Moroccan Arabic tweets.

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, and proposed a novel features representation

We have implemented:

DL algorithms: CNN, LSTM,

Classic algorithms: SVM, LR.

We performed several scenarios using several parameters:

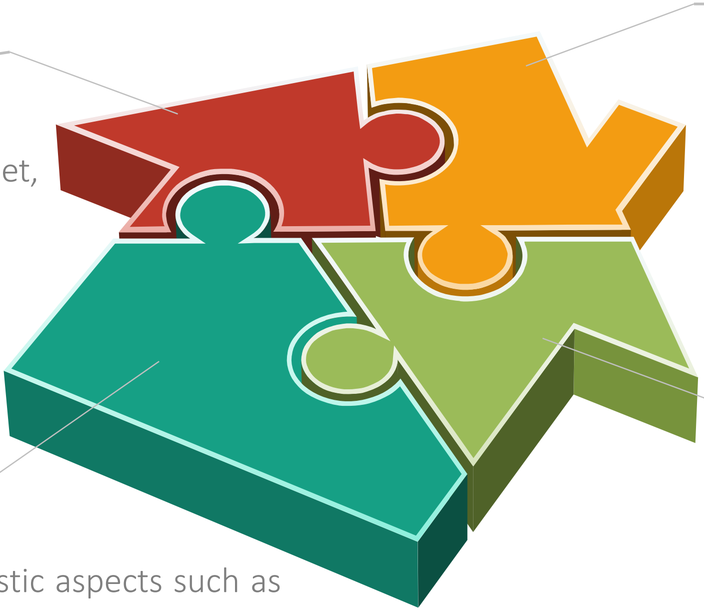
Stopwords removal

Our system achieves convincing results.

The system achieves an average precision of 91% 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

ulco Université Littoral Côte d'Opale

THE INTERNATIONAL CONFERENCE ICI2C ON INFORMATION, COMMUNICATION & CYBERSECURITY

Université Sidi Mohamed Ben Abdellah

# Towards a new Lexicon-Based features vector for Sentiment Analysis: Application to Moroccan Arabic tweets

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*