Dynamic construction of dictionaries for sentiment classification

Hanen Ameur*       Salma Jamoussi #

*,# Multimedia, InfoRmation systems and Advanced Computing Laboratory MIRACL-Sfax University, Sfax-Tunisia

IEEE International Conference on Data Mining - SENTIRE

(ICDMW 2013)
Outline

1- Introduction
2- Related work
3- Proposed method
   3.1- Acquisition and preparation of the corpus
   3.2- Dictionaries construction
   3.3- Comment classification
   3.4- Evaluation and discussion
4- Conclusion and future work
Introduction

With the emergence of **web 2.0** and the advent of **community sites**, hundreds of thousands of sentiments are shared and circulated every day on the canvas.

The sentiments have become **numerically** documented.

It is very important to process the sentiments in **many fields** (commercial, political and others).

**How can I analyze and classify automatically an opinion text based on its textual content?**
**Goal:** Determine the polarity "sentimental orientation" of a text bearer sentiment and the intensity of its polarity “valence”.

Differentiation between **positive** and **negative** sentiment.
Linguistic approach is depicted in two essential tasks:

1. Sentiment lexicon construction.
2. Subjective sentence classification.

- Manual method (by experts), e.g. [1] and [2].
  - Dictionaries-based method, e.g. [3] and [4]
  - Corpora based method, e.g. [5] and [6]
  - Hybrid method, e.g. [7]
  - Concept based method, e.g. [8], [9] and [10]

→ Count the number of positive words and the number of negative words present in a sentence. [11] and [12]
Proposed method

Acquisition and preparation of the corpus

Facebook Pages

Pretreated corpus

Learning corpus

Test corpus

Dictionaries construction

Comment classification

Utilization

Enrichment

Test corpus classified by the expert

Test corpus classified by the system

Evaluation
Proposed method

- The raw material of a sentiment classification system.
- It contains comments collected from the Facebook social network.

A lack of corpus Facebook

Automatic construction of corpus Facebook.

Acquisition and preparation of the corpus

API facebook

32 political pages

acquisition and structuring program

```xml
<?xml version="1.0" encoding="UTF-8"?>
<PageFacebook>
  <PageName/>
  <publication_id/>
  <type/>
  <Name/>
  <Link/>
  <description/>
  <message/>
  <created_time/>
  <updated_time/>
  <count_of_shares/>
  <count_of_likes/>
  <user_like/>
  <likes/>
  <comments_count/>
  <comment/>
  <message_comment/>
  <created_time_comment/>
  <like_count/>
</PageFacebook>
```
Proposed method

Acquisition and preparation of the corpus

1- Normalization

2- Filtering

3- Translation

4- Lemmatization

5- Stop words removal

Lengthening
Proposed method

Acquisition and preparation of the corpus

Facebook Pages

Learning corpus

Pretreated corpus

Test corpus

Dictionaries construction

Utilization

Enrichment

Comment classification

Test corpus classified by the expert

Test corpus classified by the system

Evaluation
Proposed method

**Goal:** Generate two dictionaries (*positive* and *negative*) covering the majority of the sentiment lexicon related to our application.
**Proposed method**

**Initial dictionaries construction**

- It is essentially based on the emotion symbols present in a comment.

<table>
<thead>
<tr>
<th>Positive emotion symbols</th>
<th>Negative emotion symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>:) :p :-* &lt;3 xD</td>
<td>:( ( &gt;:( ::-/ (^^^^)</td>
</tr>
<tr>
<td>mdr haha lol</td>
<td>pf pfet</td>
</tr>
</tbody>
</table>

**Hypothesis:**

- Emotion symbols reflect the sentiment expressed by the words that precede them in the comments.
Proposed method

Initial dictionaries construction

Principle:

- Calculate the positive and negative valences of each lexicon word.

\[
\text{valence}(m)_{\text{pol}} = \frac{\text{frequency}(m)_{\text{pol}}}{\sum_{i=0}^{n} \text{frequency}(m_i)_{\text{pol}}} \times 1000
\]

Frequency of the word \( m \) with emotion symbols having the polarity \( \text{pol} \), with \( \text{pol} = \{\text{pos}, \text{neg}\} \).

The sum of the frequencies of all words present in the dictionary of polarity \( \text{pol} \).

Each word can be present in the two dictionaries, but with different valences.
**Proposed method**

**Initial dictionaries construction**

**Principle:**
- Calculate the positive and negative valences of each lexicon word.

\[
\text{valence}(m)_{\text{pol}} = \frac{\text{frequency}(m)_{\text{pol}}}{\sum_{i=0}^{n} \text{frequency}(m_i)_{\text{pol}}} \times 1000
\]

- Compare the two positive and negative valences of the word.

<table>
<thead>
<tr>
<th>Word</th>
<th>Positive valence</th>
<th>Negative valence</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>paix (peace)</td>
<td>4.667</td>
<td>3.364</td>
<td>Positive</td>
</tr>
<tr>
<td>bonne (good)</td>
<td>1.360</td>
<td>0.509</td>
<td>Positive</td>
</tr>
<tr>
<td>triste (sad)</td>
<td>0.146</td>
<td>1.019</td>
<td>Negative</td>
</tr>
<tr>
<td>guerre (war)</td>
<td>0.236</td>
<td>0.815</td>
<td>Negative</td>
</tr>
</tbody>
</table>
**Proposed method**

**Dictionaries enrichment**

**Goals:**

- **Settle** and **adjust** the valences of words present in the dictionaries.
- **Add** new words to the dictionaries.

It is based on the words present in the **initial dictionaries** whose their valences are known.

This gentleman deserves the respect, he is the best

---

<table>
<thead>
<tr>
<th>Positive dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Words</strong></td>
</tr>
<tr>
<td>monsieur</td>
</tr>
<tr>
<td>mériter</td>
</tr>
<tr>
<td>meilleur</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Words</strong></td>
</tr>
<tr>
<td>monsieur</td>
</tr>
<tr>
<td>mériter</td>
</tr>
<tr>
<td>meilleur</td>
</tr>
</tbody>
</table>
Proposed method

**Goals:**
- Settle and adjust the valences of words present in the dictionaries.
- Add new words to the dictionaries.

It is based on the words present in the **initial dictionaries** whose their valences are known.

This gentleman deserves the respect, he is the best

**Pretreatment**

<table>
<thead>
<tr>
<th>monsieur</th>
<th>mériter</th>
<th>respect</th>
<th>meilleur</th>
</tr>
</thead>
</table>

**Valences**

\[
\text{Valence}(C)_{pol} = \frac{\sum_{i=0}^{x} \text{frequency}(m_{i})_{pol}}{\left[ \sum_{l=0}^{y} \text{frequency}(m_{l})_{pol} \right]_k} \times 1000
\]

1- Calculate the positive and negative valences of the comment.
Proposed method

**Dictionaries enrichment**

**Goals:**
- **Settle** and **adjust** the valences of words present in the dictionaries.
- **Add** new words to the dictionaries.

It is based on the words present in the **initial dictionaries** whose their valences are **known**.

This gentleman deserves the respect, he is the best

2- Compare the positive and negative valences of the comment.

Calculate the percentage of the polarity of the comment.

\[
\text{Percentage}_{pol}(C) = \frac{\text{valence}(C)_{pol}}{\text{valence}(C)_{pos} + \text{valence}(C)_{neg}}
\]
Proposed method

**Dictionaries enrichment**

- **Settle** and **adjust** the valences of words present in the dictionaries.
- **Add** new words to the dictionaries.

It is based on the words present in the initial dictionaries whose their valences are known.

This gentleman deserves the respect, he is the best

---

**Goals:**

1. **Settle** and **adjust** the valences of words present in the dictionaries.
2. **Add** new words to the dictionaries.

**Valences**

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequencies</th>
<th>Valences</th>
</tr>
</thead>
<tbody>
<tr>
<td>monsieur</td>
<td>12.5258, 13.2359</td>
<td>-0.2879, 0.3042</td>
</tr>
<tr>
<td>mériter</td>
<td>81.3735, 82.0836</td>
<td>1.8664, 1.8866</td>
</tr>
<tr>
<td>meilleur</td>
<td>241.4172, 242.1273</td>
<td>-5.5568, -5.5649</td>
</tr>
<tr>
<td>respect</td>
<td>0.7107</td>
<td>0.0163</td>
</tr>
</tbody>
</table>

**Pretreatment**

Modify the valences of the **existent** words

Add **nonexistent** words

---

%Pos = 0.7107
# Proposed method

## Dictionaries enrichment

Apply the enrichment principle on **30000 comments**.

### Quantitative viewpoint

<table>
<thead>
<tr>
<th></th>
<th>Initial dictionaries</th>
<th>Enriched dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td><strong>23705</strong></td>
<td><strong>5106</strong></td>
<td><strong>44302</strong></td>
</tr>
</tbody>
</table>

### Qualitative viewpoint

<table>
<thead>
<tr>
<th>Word</th>
<th>Initial dictionaries</th>
<th>Enriched dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive valence</td>
<td>Negative valence</td>
</tr>
<tr>
<td>peace</td>
<td>4.6702</td>
<td>3.3646</td>
</tr>
<tr>
<td>love</td>
<td>1.2491</td>
<td>1.3254</td>
</tr>
<tr>
<td>urgency</td>
<td>0.1012</td>
<td>0</td>
</tr>
<tr>
<td>educe</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Adjust** the valences of existing words.
- **Correct** the valences of existing words.
- **Add** words to the dictionaries.
**Proposed method**

**Handling negation**

<table>
<thead>
<tr>
<th>The negation particles</th>
</tr>
</thead>
<tbody>
<tr>
<td>'ne', 'n', 'pas', 'ni', 'jamais', 'aucun', 'no', 'none', 'not', 'neither', 'never', 'ever', 'لم', 'لن', 'لا', 'ﻻ'</td>
</tr>
</tbody>
</table>

- **Reverse the polarity** of all words **preceded** by one of the negation particles.

- **At the enrichment dictionaries**

  **First method:**

  Add words preceded by a negation particle to the dictionary that corresponds to the **inverse comment’s polarity**.

  **Second method:**

  Reverse also the **frequencies** of all words preceded by a negation particle, when calculating the valences of the comment.
Proposed method

Acquisition and preparation of the corpus

Pretreated corpus

Learning corpus

Test corpus

Dictionaries construction

Comment classification

Utilization

Enrichment

Test corpus classified by the expert

Test corpus classified by the system

Evaluation

Acquisition and preparation of the corpus

Facebook Pages

Acquisition and preparation of the corpus

Pretreated corpus

Learning corpus

Test corpus

Dictionaries construction

Comment classification

Utilization

Enrichment

Test corpus classified by the expert

Test corpus classified by the system

Evaluation
**Goal:** Determine the **polarities of comments** (positive/negative) using the dictionaries obtained in the previous step.

We love you Mister President Obama.

**Pretreatment**

<table>
<thead>
<tr>
<th>Word</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td>aimer</td>
<td>2.8696</td>
</tr>
<tr>
<td>monsieur</td>
<td>0.5626</td>
</tr>
<tr>
<td>président</td>
<td>4.3439</td>
</tr>
<tr>
<td>obama</td>
<td>5.8631</td>
</tr>
</tbody>
</table>

**Valences**

<table>
<thead>
<tr>
<th>Valence</th>
<th>aimer</th>
<th>monsieur</th>
<th>président</th>
<th>obama</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.2039</td>
<td>0.3058</td>
<td>2.3450</td>
<td>2.1411</td>
</tr>
<tr>
<td>Negative</td>
<td>2.8696</td>
<td>0.5626</td>
<td>4.3439</td>
<td>5.8631</td>
</tr>
</tbody>
</table>

\[ \sum P_{com} = 10.207 \]

\[ N_{com} = 4.4861 \]
Proposed method

Acquisition and preparation of the corpus

Pretreated corpus

Learning corpus

Test corpus

Facebook Pages

Dictionaries construction

Utilization

Enrichment

Comment classification

Test corpus classified by the expert

Test corpus classified by the system

Evaluation

Acquisition and preparation of the corpus

Facebook Pages

Dictionaries construction

Utilization

Enrichment

Comment classification

Test corpus classified by the expert

Test corpus classified by the system

Evaluation

Acquisition and preparation of the corpus

Facebook Pages

Dictionaries construction

Utilization

Enrichment

Comment classification

Test corpus classified by the expert

Test corpus classified by the system

Evaluation

Acquisition and preparation of the corpus

Facebook Pages

Dictionaries construction

Utilization

Enrichment

Comment classification

Test corpus classified by the expert

Test corpus classified by the system

Evaluation
The external evaluation

Measure the adequacy between the results of our system classification and that made by the expert.

<table>
<thead>
<tr>
<th>Political dictionaries</th>
<th>Initial</th>
<th>Enriched</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w/ negation</td>
<td>w/o negation</td>
</tr>
<tr>
<td>Error rate</td>
<td>35.02</td>
<td>33.75</td>
</tr>
<tr>
<td>Recall</td>
<td>65.08</td>
<td>66.39</td>
</tr>
<tr>
<td>Accuracy</td>
<td>66.06</td>
<td>66.52</td>
</tr>
<tr>
<td>F-score</td>
<td>65.57</td>
<td>66.45</td>
</tr>
</tbody>
</table>
Conclusion and future work

- Realize an **automatic classification** system of comments derived from the social network **Facebook**.

- Construction of **dictionaries** covering the majority of the sentiment lexicon from our learning corpus (based on **linguistic approach**).

  They are used for the calculation of the positive and negative **polarities of comments**.

- Propose a **statistic method** to construct a sentiment lexicon served for the sentiment classification.

- Find a **vector representation** of words and sentences and exploit the classification methods (supervised, unsupervised, and semi-supervised).
References


References


Thank you :)