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

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.

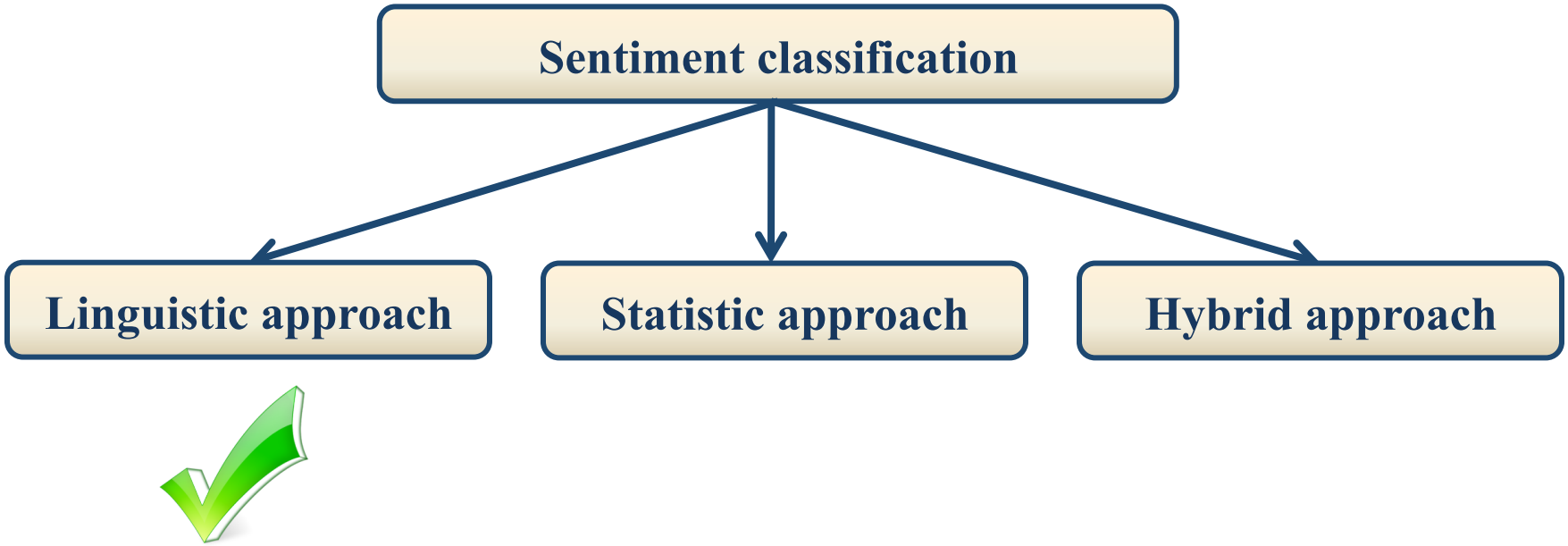
It is very important to process the sentiments in **many fields** (commercial, political and others).

How can I analyze and classify automatically an opinion text based on its textual content?

Introduction

Goal: Determine the polarity "**sentimental orientation**" of a text bearer sentiment and the intensity of its polarity "**valence**".

Differentiation between **positive** and **negative** sentiment.



Related work (Linguistic approach)

- **Linguistic approach** is depicted in **two** essential tasks:

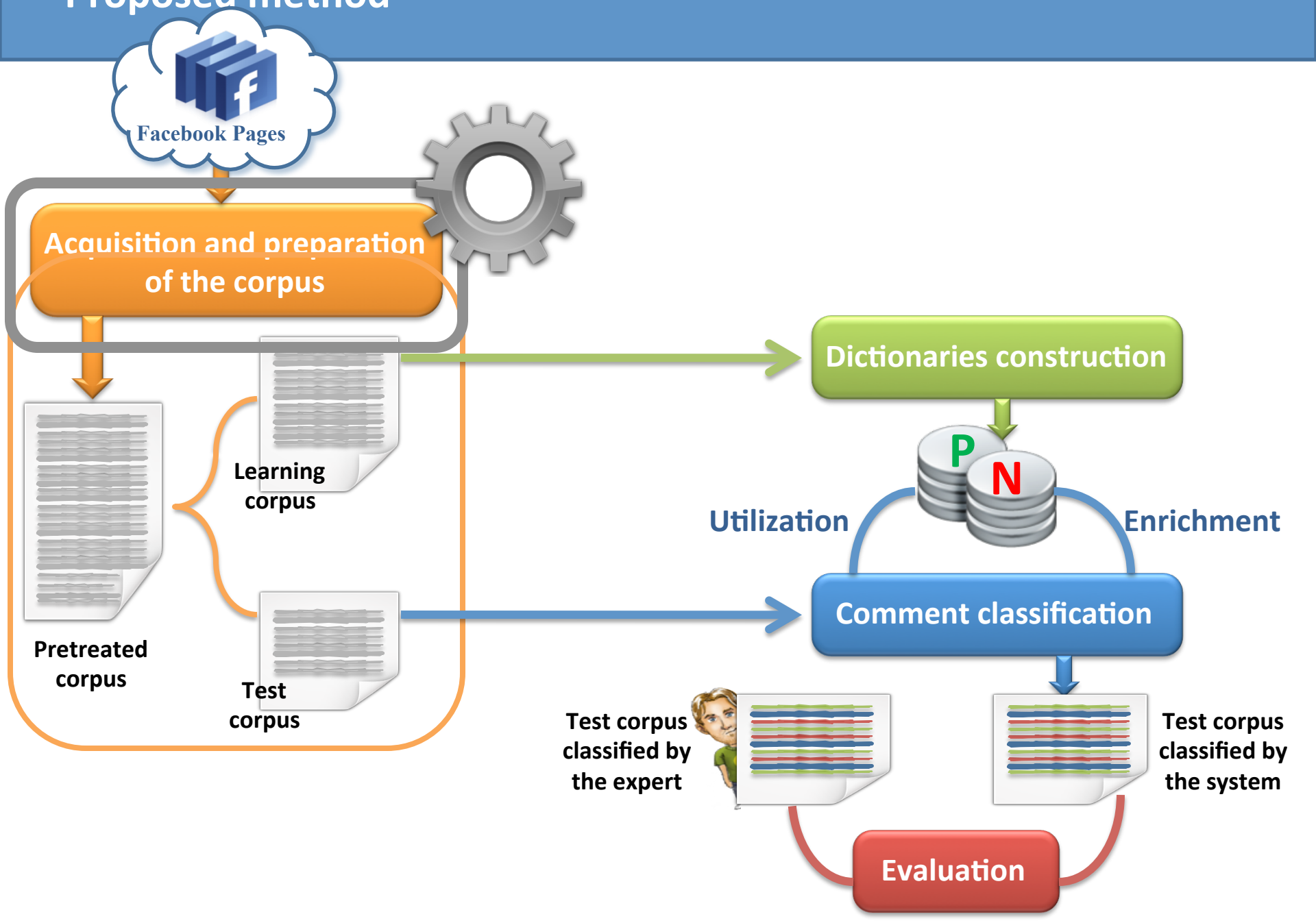
  1-Sentiment lexicon construction.

  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



Proposed method

Acquisition and preparation of the corpus

- 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.

32 political pages

API facebook

acquisition and structuring program

```
<?xml version="1.0" encoding="UTF-8"?>
<PagesFacebook>
  <Page Name="..." likes="..." talking_about_count="...">
    <Publication id="...">
      <Type>...</Type>
      <Name>...</Name>
      <Link>...</Link>
      <Description>...</Description>
      <Message>...</Message>
      <created_time>...</created_time>
      <updated_time>...</updated_time>
      <Count_of_shares>...</Count_of_shares>
      <Likes count="...">
        <User_like>...</User_like>
      </Likes>
      <Comments count="...">
        <Comment>
          <User_comment>...</User_comment>
          <message_comment>...message_comment>
          <created_time_comment>...</created_time_comment>
          <like_count>...</like_count>
        </Comment>
      </Comments>
    </Publication>
  </Page>
</PagesFacebook>
```

Proposed method

Acquisition and preparation of the corpus

@ Ridha I am against the frauds committed with Ben Ali -- thief :(((http://www.facebook.com/.

1-Normalization

~~@ Ridha~~ i am against the frauds committed with ben ali thief :(((~~http://www.facebook.com/.~~

2-Filtering

i am against the frauds committed with ben ali thief :(((

Google translate **3-Translation**

je suis contre les fraudes commises avec ben ali voleur :(((

4-Lemmatization

~~je être~~ contre ~~le~~ fraude commettre ~~avec~~ ben ali voleur :(((

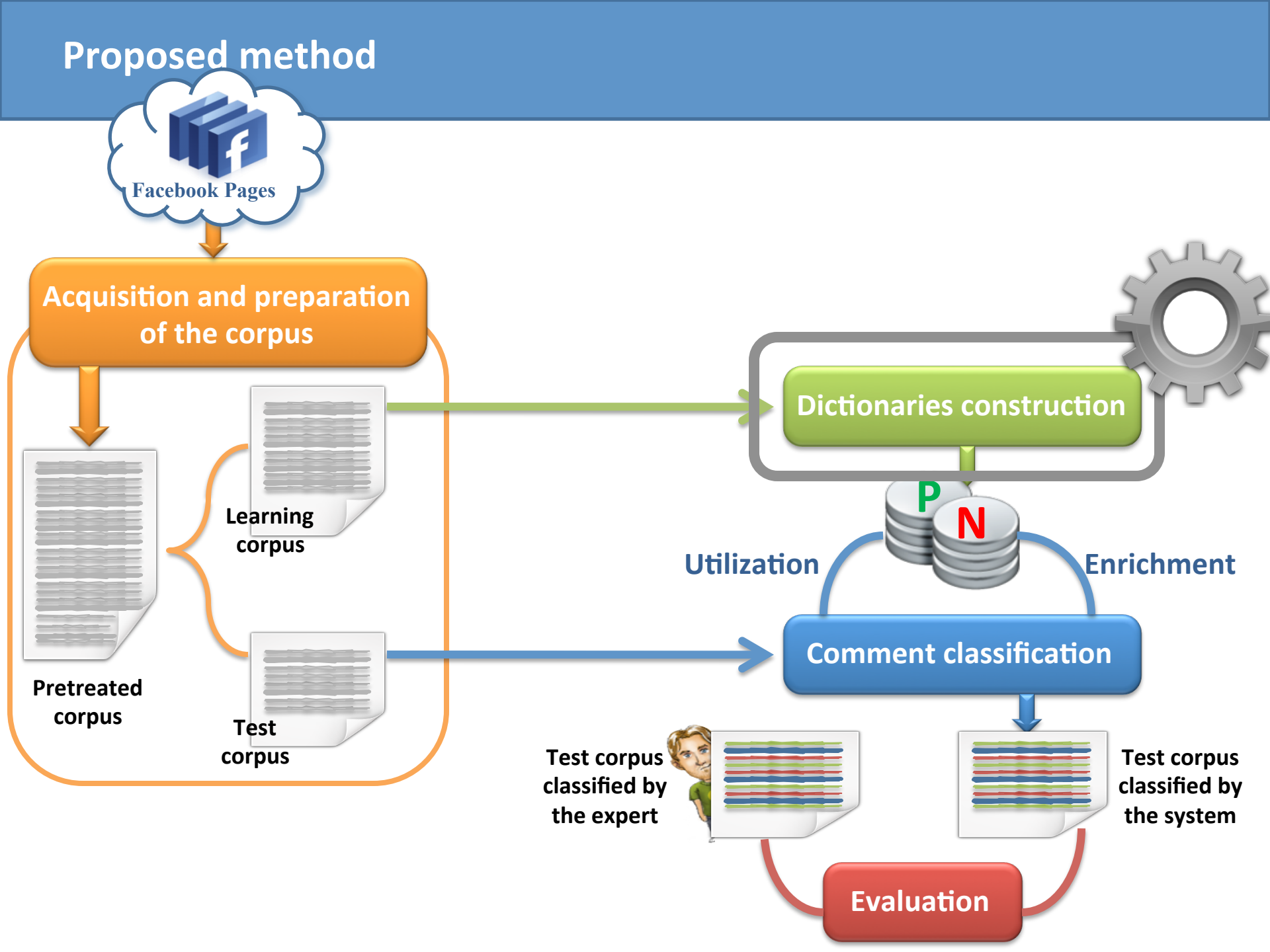
5-Stop words removal

contre fraude commettre ben ali voleur :(((

Lengthening

contre fraude commettre ben ali voleur :(

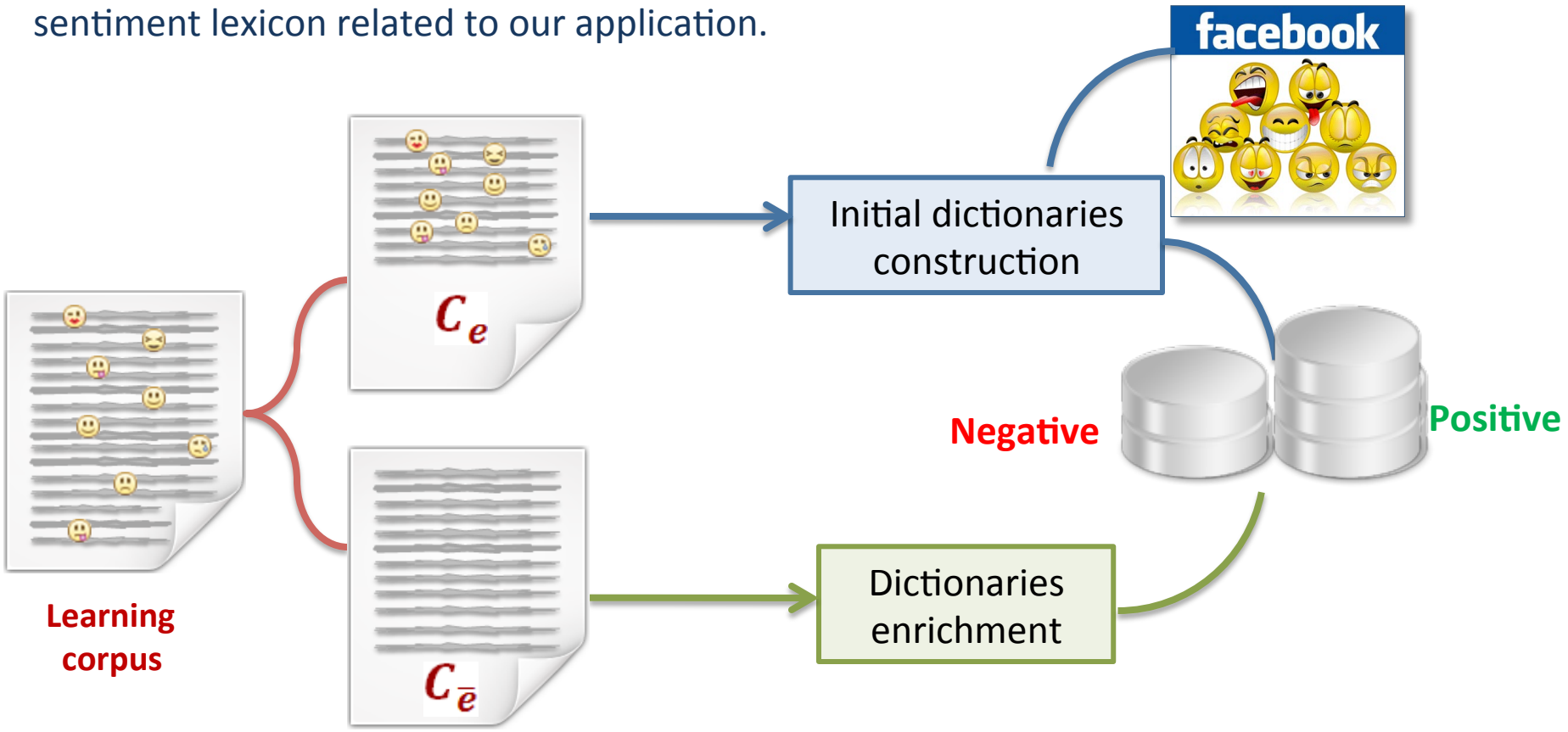
Proposed method



Proposed method

Dictionaries construction

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



Proposed method

Dictionaries construction

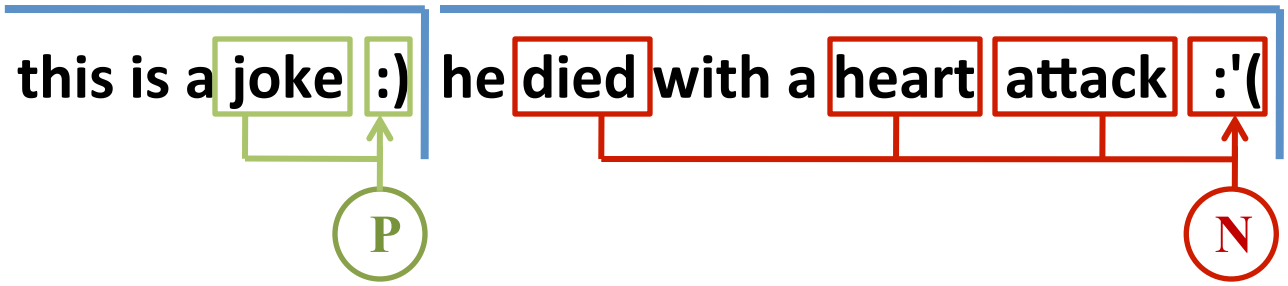
Initial dictionaries construction

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

Positive emotion symbols	Negative emotion symbols
:) :p :-* <3 xD	:(:'(>:(:-/ (^^^)
mdr haha lol عه	pf pfet بفط

Hypothesis :

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



Proposed method

Dictionaries construction

Initial dictionaries construction

Principle :

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

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

Frequency of the **word m** with emotion symbols having the polarity **pol**,
The sum of the frequencies of **all words** present in the **dictionary** of polarity **pol**.
with **pol={pos,neg}.**

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

Proposed method

Dictionaries construction

Initial dictionaries construction

Principle :

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

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

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

Word	Positive valence	Negative valence	Polarity
paix (peace)	4.667	3.364	Positive
bonne (good)	1.360	0.509	Positive
triste (sad)	0.146	1.019	Negative
guerre (war)	0.236	0.815	Negative

Proposed method

Dictionaries construction

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

Pretreatment

monsieur | mériter | respect | meilleur

?

Positive dictionary

Words	Frequencies	Valences
monsieur	12.5258	0.2879
mériter	81.3735	1.8664
⋮		
meilleur	241.4172	5.5568

Negative dictionary

Words	Frequencies	Valences
monsieur	14.4682	0.4608
mériter	16.1875	0.5236
⋮		
meilleur	67.7028	2.3637

Proposed method

Dictionaries construction

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

Pretreatment

monsieur | mériter | respect | meilleur

Valences

7.3317

2.9928

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

$$valence(C)_{pol} = \frac{\sum_{i=0}^{x} frequency(m_i)_{pol}}{[\sum_{l=0}^{y} frequency(m_l)_{pol}]_k} \times 1000$$

Proposed method

Dictionaries construction

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

Pretreatment

monsieur | mériter | respect | meilleur

Valences

7.3317

2.9928

%Pos= 0.7107

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

Calculate the percentage of the polarity of the comment.

$$Pourcentage_{pol}(C) = \frac{valence(C)_{pol}}{valence(C)_{pos} + valence(C)_{neg}}$$

Proposed method

Dictionaries construction

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

Pretreatment

monsieur | mériter | respect | meilleur

Valences

7.3317

2.9928

%Pos= 0.7107

Positive dictionary

Words	Frequencies	Valences
monsieur	~~12.5258~~ 13.2359	~~0.2879~~ 0.3042
mériter	~~81.3735~~ 82.0836	~~1.8664~~ 1.8866
⋮		
meilleur	~~241.4172~~ 242.1273	~~5.5568~~ 5.5649
respect	0.7107	0.0163

Modify the valences of the **existent** words

Add **nonexistent** words

Proposed method

Dictionaries construction

Dictionaries enrichment

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

•Quantitative viewpoint

Initial dictionaries		Enriched dictionaries	
Positive	Negative	Positive	Negative
23705	5106	44302	46420

•Qualitative viewpoint

	Initial dictionaries		Enriched dictionaries	
	Positive valence	Negative valence	Positive valence	Negative valence
peace	4.6702	3.3646	5.5566	2.5632
love	1.2491	1.3254	1.1789	0.5572
urgency	0.1012	0	0.0686	0.0988
educe	0	0	0	0.213

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

Proposed method

Dictionaries construction

Handling negation

The negation particles
'ne', 'n', 'pas', 'ni', 'jamais', 'aucun', 'no‘, 'none', 'not', 'neither', 'never', 'ever', 'لا', 'لن', 'لم'

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

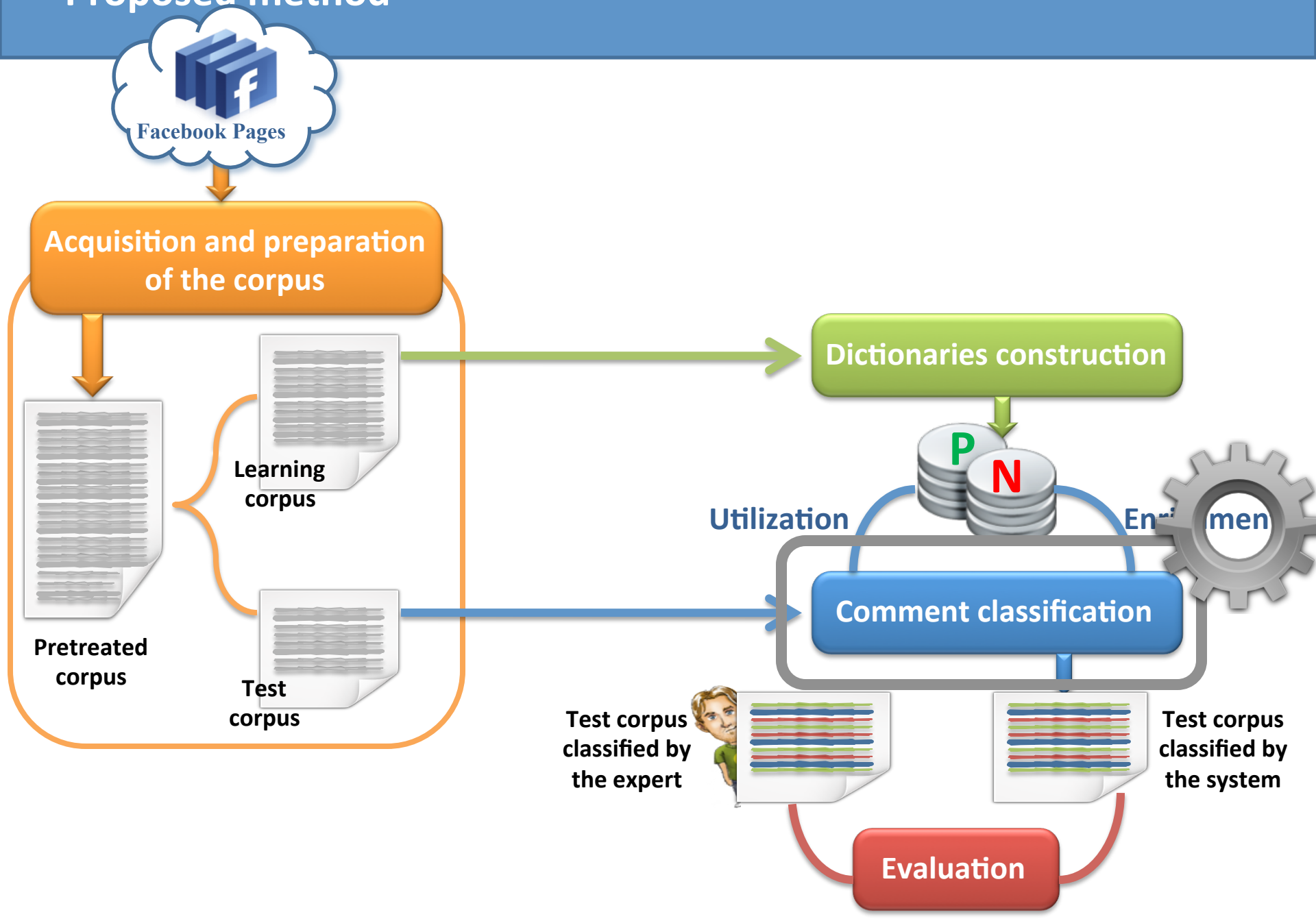
First method :

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



Proposed method

Comment classification

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

We love you Mister President Obama.

Pretreatment

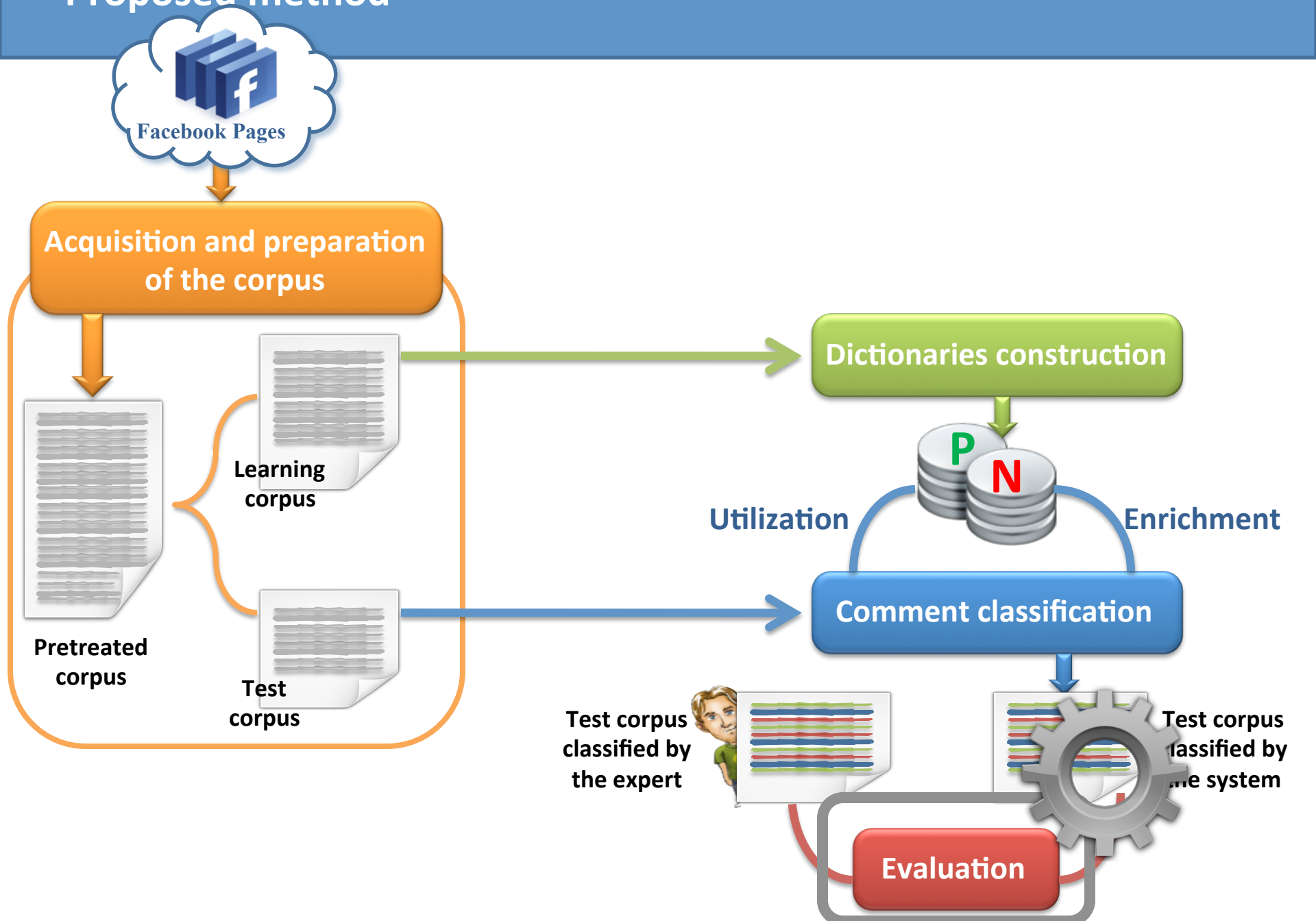
aimer	monsieur	président	obama
2.8696	0.5626	4.3439	5.8631
0.2039	0.3058	2.3450	2.1411

Valences

P_{com} : 10.207

N_{com} : 4.4861

Proposed method



Proposed method

Evaluation

The external evaluation

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

Political dictionaries	Initial		Enriched	
	w/ negation	w/o negation	Method 1	Method 2
Error rate	35.02	33.75	18.75	19.83
Recall	65.08	66.39	80.76	80.76
Accuracy	66.06	66.52	81.01	80.07
F-score	65.57	66.45	80.89	80.42

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

[1] Nasukawa, T., Yi, J., Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the Conference on Knowledge Capture (K-CAP). New York, NY, USA (2003)

[2] Wilson, T., Wiebe, J., Homann, P., Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). Vancouver, CA (2005)

[3] J. Kamps and M. Marx, "Words with attitude," in 1st International WordNet Conference, Mysore, India, 2002.

[4] M. Taboada, C. Anthony, and K. Voll, "Methods for creating semantic orientation dictionaries," in Conference on Language Resources and Evaluation (LREC), 2006.

[5] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia, July 2006.

[6] X. Ding and B. Liu, "The utility of linguistic rules in opinion mining," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2007.

References

[7] A. Pak and P. Paroubek, “Construction dun lexique affectif pour le franais partir de twitter,” TALN 2010. Btiment 508, F-91405 Orsay Cedex, France: Universit de Paris-Sud, juillet2010.

[8] E. Cambria, T. Mazzocco, and A. Hussain, “Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining,” Biologically Inspired Cognitive Architectures, vol. 4, p. 4153, 2013.

[9] M. Grassi, E. Cambria, A. Hussain, and F. Piazza, “Sentic web: A new paradigm for managing social media affective information,” Cognitive Computation, vol. 3, no. 3, pp. 480– 489, 2011.

[10] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” IEEE Intelligent Systems, vol. 28, no. 2, pp. 15–21, 2013.

[11] M. Hu and B. Liu, “Mining opinion features in customer reviews,” in Proceedings of AAAI, Seattle, WA, USA, 2004.

[12] D. Poirier, F. Fessant, C. Bothorel, E. Guimier De Neef, and M. Boull´e, “Approches statistique et linguistique pour la classification de textes d’opinion portant sur les films,” vol. E17, pp. 147–169, 2009.

Thank you :)