

French Social Media Mining : Expertise and Sentiment

Amine Abdaoui

▶ **To cite this version:**

Amine Abdaoui. French Social Media Mining : Expertise and Sentiment. Artificial Intelligence [cs.AI]. Université Montpellier, 2016. English. NNT : 2016MONTT249 . tel-01507494v2

HAL Id: tel-01507494

https://hal-lirmm.ccsd.cnrs.fr/tel-01507494v2

Submitted on 25 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESIS

To obtain the degree of Doctor of Philosophy (Ph.D.)

Awarded by **University of Montpellier**

Prepared at **I2S*** Graduate School,
LIRMM Research Unit, ADVANSE Team.

Speciality: **Computer Science**

Defended by **Amine ABDAOUI**
abdaoui@lirmm.fr

French Social Media Mining: Expertise and Sentiment

Defended on 05/12/2016 in front of a jury composed by:

Andrea Tagarelli	HDR	University of Calabria	Reviewer
Julien Velcin	HDR	University of Lyon 2	Reviewer
Alexandre Allauzen	HDR	University of Paris Sud	Examiner
Philippe Lenca	PU	Telecom Bretagne	Examiner
Pascal Poncelet	PU	University of Montpellier	President
Jérôme Azé	PU	University of Montpellier	Director
Sandra Bringay	PU	Montpellier Paul Valery	Co-Director

One may walk over the highest mountain one step at a time

* **I2S:** Information, Structures and Systems.

Abstract

Social Media has changed the way we communicate between individuals, within organizations and communities. The availability of these social data opens new opportunities to understand and influence the user behavior. Therefore, Social Media Mining is experiencing a growing interest in various scientific and economic circles. In this thesis, we are specifically interested in the users of these networks whom we try to characterize in two ways: (i) their expertise and their reputations and (ii) the sentiments they express.

Conventionally, social data is often mined according to its network structure. However, the textual content of the exchanged messages may reveal additional knowledge that can not be known through the analysis of the structure. Until recently, the majority of work done for the analysis of the textual content was proposed for English. The originality of this thesis is to develop methods and resources based on the textual content of the messages for French Social Media Mining.

In the first axis, we initially suggest to predict the user expertise. For this, we used forums that recruit health experts to learn classification models that serve to identify messages posted by experts in any other health forum. We demonstrate that models learned on appropriate forums can be used effectively on other forums. Then, in a second step, we focus on the user reputation in these forums. The idea is to seek expressions of trust and distrust expressed in the textual content of the exchanged messages, to search the recipients of these messages and use this information to deduce users' reputation. We propose a new reputation measure that weighs the score of each response by the reputation of its author. Automatic and manual evaluations have demonstrated the effectiveness of the proposed approach.

In the second axis, we focus on the extraction of sentiments (emotions and polarity). For this, we started by building a French lexicon of sentiments and emotions that we call FEEL (French Expanded Emotions Lexicon). This lexicon is built semi-automatically by translating and expanding its English counterpart NRC EmoLex. We then compare FEEL with existing French lexicons from literature on reference benchmarks. The results show that FEEL improves the classification of French texts according to their polarities and emotions. Finally, we propose to evaluate different features, methods and resources for the classification of sentiments in French. The conducted experiments have identified useful features and methods in the classification of sentiments for different types of texts. The learned systems have been particularly efficient on reference benchmarks.

Generally, this work opens promising perspectives on various analytical tasks of Social Media Mining including: (i) combining multiple sources in mining Social Media users; (ii) multi-modal Social Media Mining using not just text but also image, videos, location, etc. and (iii) multilingual sentiment analysis.

Résumé

Les médias sociaux ont changé notre manière de communiquer entre individus, au sein des organisations et des communautés. La disponibilité de ces données sociales ouvre de nouvelles opportunités pour comprendre et influencer le comportement des utilisateurs. De ce fait, la fouille des médias sociaux connait un intérêt croissant dans divers milieux scientifiques et économiques. Dans cette thèse, nous nous intéressons spécifiquement aux utilisateurs de ces réseaux et cherchons à les caractériser selon deux axes : (i) leur expertise et leur réputation et (ii) les sentiments qu'ils expriment.

De manière classique, les données sociales sont souvent fouillées selon leur structure en réseau. Cependant, le contenu textuel des messages échangés peut faire émerger des connaissances complémentaires qui ne peuvent être connues via la seule analyse de la structure. Jusqu'à récemment, la majorité des travaux concernant l'analyse du contenu textuel était proposée pour l'Anglais. L'originalité de cette thèse est de développer des méthodes et des ressources basées sur le contenu pour la fouille des réseaux sociaux pour la langue Française.

Dans le premier axe, nous proposons d'abord d'identifier l'expertise des utilisateurs. Pour cela, nous avons utilisé des forums qui recrutent des experts en santé pour apprendre des modèles de classification qui servent à identifier les messages postés par les experts dans n'importe quel autre forum. Nous démontrons que les modèles appris sur des forums appropriés peuvent être utilisés efficacement sur d'autres forums. Puis, dans un second temps, nous nous intéressons à la réputation des utilisateurs dans ces forums. L'idée est de rechercher les expressions de confiance et de méfiance exprimées dans les messages, de rechercher les destinataires de ces messages et d'utiliser ces informations pour en déduire la réputation des utilisateurs. Nous proposons une nouvelle mesure de réputation qui permet de pondérer le score de chaque réponse selon la réputation de son auteur. Des évaluations automatiques et manuelles ont démontré l'efficacité de l'approche.

Dans le deuxième axe, nous nous sommes focalisés sur l'extraction de sentiments (polarité et émotion). Pour cela, dans un premier temps, nous avons commencé par construire un lexique de sentiments et d'émotions pour le Français que nous appelons FEEL (French Expanded Emotion Lexicon). Ce lexique est construit de manière semi-automatique en traduisant et en étendant son homologue Anglais NRC EmoLex. Nous avons ensuite comparé FEEL avec les lexiques Français de la littérature sur des benchmarks de référence. Les résultats ont montré que FEEL permet d'améliorer la classification des textes Français selon leurs polarités et émotions.

Dans un deuxième temps, nous avons proposé d'évaluer de manière assez exhaustive différentes méthodes et ressources pour la classification de sentiments en Français. Les expérimentations menées ont permis de déterminer les caractéristiques utiles dans la classification de sentiments pour différents types de textes. Les systèmes appris se sont montrés particulièrement efficaces sur des benchmarks de référence.

De manière générale, ces travaux ont ouvert des perspectives prometteuses sur diverses tâches d'analyse des réseaux sociaux pour la langue Française incluant: (i) combiner plusieurs sources pour transférer la connaissance sur les utilisateurs des réseaux sociaux; (ii) la fouille des réseaux sociaux en utilisant les images, les vidéos, les géolocalisations, etc. et (iii) l'analyse multilingues de sentiment.

Acknowledgements

This manuscript describes the scientific outputs of my thesis but it does not relate the good days spent with the great people to whom I will always still grateful for making this happen. In this short Acknowledgement section, I will try to briefly thank them for being here.

First of all, I would like to thank my supervisors (Professor Jérome Azé and Professor Sandra Bringay) for their encouragement. Working with you during three years was an exciting advanture. Then, I have to express my gratitude to all the members of the jury for their interesting questions and stimulating discussions. A special acknowledgment goes to the very special Professor Pascal Poncelet.

During this three years, I enjoyed meeting, working and lauphing with particularly talented people. I want to thank all my friends and collegues for sharing these moments: Mike, Vijay, Antonio, Sarah, Samiha, Eric, Jessica and Lynda. I can not forget my wife, my parrents and my sister for their lovely support.

Finally, I am grateful to the Algerian Ministry of Higher Education and Scientific Research for giving the fundings related to this thesis.

Contents

1 Introduction **1**

1.1 Context and Motivations 2
1.2 Research Contributions 6
1.2.1 Data Collection and Annotation 6
1.2.2 Methods for User Expertise and Reputation 6
1.2.3 French Sentiment Lexicon 7
1.2.4 Sentiment Classification Process 7
1.3 Thesis Organization 7
1.3.1 Part I: User Expertise and Reputation 7
1.3.2 Part II: Sentiment Analysis 9
1.4 Publications 9
1.4.1 International Journals (2) 9
1.4.2 International Conferences (6) 10
1.4.3 French Conferences (2) 10
1.4.4 Workshops (2) 10

2 Social Media Mining: State of the Art **13**

2.1 Introduction 14
2.2 Network Analysis in Social Media 14
2.2.1 Extracting Social Networks 15
2.2.2 Mining the Expertise in Social Networks 17
2.3 Text Mining in Social Media 18
2.3.1 Linguistic Pre-processing 19
2.3.2 Text Representation 20
2.3.3 Text Categorization 21
2.4 Sentiment Analysis 25
2.4.1 Lexicons 26
2.4.2 Classification Methods 30
2.4.3 Benchmarks 33
2.5 Conclusion 35

I Expertise and Reputation 37

3 Predicting User Expertise 39
3.1 Introduction . . . 40
3.2 Materials and Methods . . . 41
3.2.1 Corpora . . . 41
3.2.2 Pre-processings . . . 42
3.2.3 Annotations . . . 43
3.2.4 Classification . . . 44
3.3 Experiments . . . 45
3.3.1 Cross Validation . . . 45
3.3.2 Training and Testing on Different Data sets . . . 46
3.4 Discussions . . . 47
3.4.1 Results Interpretation . . . 47
3.4.2 Error Analysis . . . 48
3.5 Conclusion . . . 49

4 Computing User Reputation 51
4.1 Introduction . . . 52
4.2 Materials and Methods . . . 53
4.2.1 Theoretical Framework . . . 53
4.2.2 Corpora . . . 54
4.2.3 Extracting the Interaction Network . . . 56
4.2.4 Predicting Positive, Negative and Neutral Replies . . . 57
4.2.5 Proposed Metrics . . . 58
4.3 Evaluations and Discussions . . . 59
4.3.1 Evaluating the Network Extraction Step . . . 59
4.3.2 Evaluating the Trust Prediction Step . . . 60
4.3.3 Evaluating the Proposed Metric . . . 61
4.4 Conclusion . . . 63

II Sentiment Analysis 65

5 The FEEL Lexicon 67
5.1 Introduction . . . 68
5.2 Compilation Process . . . 68
5.2.1 Automatic Creation . . . 69
5.2.2 Validating the Translations . . . 70
5.2.3 Evaluating the Sentiments . . . 71
5.3 Comparative Study . . . 74
5.3.1 Evaluation in a Polarity Classification Task . . . 76
5.3.2 Evaluation in an Emotion Classification Task . . . 78
5.4 Conclusion . . . 80

6 French Sentiment Classification **83**
6.1 Introduction . . . 84
6.2 Features and Methods . . . 84
6.2.1 Word Ngrams . . . 85
6.2.2 Preprocessings . . . 85
6.2.3 Handling Negation . . . 85
6.2.4 Lexicon Features . . . 85
6.2.5 Syntactic Features . . . 86
6.2.6 Word Embeddings . . . 86
6.2.7 Feature Subset Selection . . . 86
6.2.8 Classifier . . . 86
6.3 Experimentations . . . 87
6.3.1 Tuning on Cross Validation . . . 87
6.3.2 Evaluating the Selected Configurations . . . 90
6.4 Conclusion . . . 94

General Conclusion and Future Work **96**

7 Conclusion and Future Work **99**
7.1 Thesis Summary . . . 100
7.1.1 Part I: User Expertise and Reputation . . . 100
7.1.2 Part II: Sentiment Analysis . . . 101
7.2 Future Research . . . 102
7.2.1 User Mining by Combining Multiple Sources . . . 102
7.2.2 Multi-modal Social Media Mining . . . 102
7.2.3 Enhancing Sentiment Analysis . . . 103
7.2.4 Ethical Considerations in Social Media Mining . . . 103

References **105**

List of Figures

1.1 The number of active users in the best-known Social Media websites as of April 2016 (source: www.statista.com). 2
1.2 The dimensions of the contextual knowledge that can be extracted from social data. 4
1.3 The organization of the remaining of this dissertation. 8

2.1 A simple social network of three users represented by an oriented graph. 15
2.2 The Text Mining process. 18
2.3 The number of text documents for each polarity class in the three benchmarks of DEFT'07. 34
2.4 The distribution of tweets in each class of the DEFT'15 benchmark. . 35

3.1 The number of documents and the average number of words per document in each corpus. 42
3.2 The number of medical concepts, emotion terms, uncertainty markers, misspellings and question marks in each corpus. 45
3.3 The percentage of posts correctly and incorrectly classified by a majority vote using 10-fold cross validation on AlloDocteurs. 49

4.1 The representation of a thread discussion using multigraph where nodes represent users (v_i) and edges represent replies (m_i). 54
4.2 The computed reputations according to the number of postings in each forum. 61

5.1 The FEEL compilation process. 69
5.2 The distribution (in a log_{10} scale) of FEEL terms in the training set of the Climate benchmark. 72
5.3 The percentage of terms in each lexicon according to their length (number of words). 74

6.1 Steps of our Feature Engineering Process. 87
6.2 The F1-measures obtained by our system compared to the maximum and average valudes obtained at DEFT'07 for each benchmark. 93

6.3 The F1-measures obtained by our system compared to the maximum and average valudes obtained at DEFT'07 for each benchmark. 93

List of Tables

1.1 A summary of the publications related to each thesis part. 11

2.1 The interpretations of Kapa values. 25
2.2 A summary of English and French Sentiment Lexicons. 29
2.3 The benchmarks used in sentiment-related French challenges. 33
2.4 The average number of words per document in each benchmark. . . . 35

3.1 The weighted F1-measures obtained with 10-fold cross validation on AlloDocteurs. 46
3.2 The weighted F1-measures obtained with 10-fold cross validation on MaSanteNet. 46
3.3 The weighted F1-measures obtained with AlloDocteurs as training set and MaSanteNet as testing set. 47
3.4 The weighted F1-measures obtained with MaSanteNet as training set and AlloDocteurs as testing set. 47

4.1 The number of users having each rank and the range of postings to acquire it in CancerDuSein.org. 55
4.2 The number of users having each rank in Forum-thyroide.net. 56
4.3 The number of annotated threads and messages from each website. . 59
4.4 The percentage of links found by one, two or three annotators in each forum. 60
4.5 The evaluation of the network extraction heuristic on both forums. . 60
4.6 The evaluation of the trust prediction on both forums. 61
4.7 The average reputation of each user rank in CancerDuSein.org. 62
4.8 The average reputation of each user rank in Forum-thyroide.net. . . . 62

5.1 The intersections between polarities and emotions in FEEL. 71
5.2 The annotators agreement for polarity and emotions (arithmetic mean) in each annotation type. We present the Fleiss' Kappa and the percentage of terms for which all annotators chose the same sentiment class. 73
5.3 The evaluation of the sentiments associated with the chosen subset of terms. 73

5.4 The intersections between the terms in each couple of lexicons. 75
5.5 The number of positive, negative and neutral terms in each lexicon. . 75
5.6 The percentage of common terms having the same polarity between each couple of lexicons. 75
5.7 The polarity classification results on the benchmark See & Read. . . . 76
5.8 The polarity classification results on Political Debate. 77
5.9 The polarity classification results on the benchmark Videos Games. . 77
5.10 The polarity classification results on the benchmark Climate Polarity. 78
5.11 The emotion classification results when considering 18 emotional classes. 79
5.12 The emotion classification results when considering 4 emotional classes. 79

6.1 The selected features and parameters by cross validation on the training set of each benchmark. 89
6.2 The obtained results after each step by 10-folds cross validation on the benchmark See & Read (3 classes). 90
6.3 The obtained results after each step by 10-folds cross validation on the benchmark Videos Games (3 classes). 91
6.4 The obtained results after each step by 10-folds cross validation on the benchmark Parliamentary Debate (2 classes). 91
6.5 The obtained results after each step by 10-folds cross validation on the benchmark Climate - Polarity (3 classes). 91
6.6 The obtained results after each step by 10-folds cross validation on the benchmark Climate - Subjectivity (4 classes). 92
6.7 The obtained results after each step by 3-folds cross validation on the benchmark Climate - Emotion (18 classes). 92
6.8 The results obtained by the selected configurations on each benchmark. 92

Chapter

1 Introduction

Contents

1.1 Context and Motivations 2

1.2 Research Contributions 6

1.2.1 Data Collection and Annotation 6

1.2.2 Methods for User Expertise and Reputation 6

1.2.3 French Sentiment Lexicon 7

1.2.4 Sentiment Classification Process 7

1.3 Thesis Organization 7

1.3.1 Part I: User Expertise and Reputation 7

1.3.2 Part II: Sentiment Analysis 9

1.4 Publications 9

1.4.1 International Journals (2) 9

1.4.2 International Conferences (6) 10

1.4.3 French Conferences (2) 10

1.4.4 Workshops (2) 10

1.1 Context and Motivations

Social Media is the group of internet-based applications that allow the creation and exchange of user-generated content (Kaplan and Haenlein, 2010). Over the past few years, many Social Media categories have emerged including: social networks (e.g. Facebook), microblogging (e.g. Twitter), online forums (e.g. StackExchange), photo sharing (e.g. Instagram), video sharing (e.g. Youtube), social gaming (World of Warcraft), etc. These web services have revolutionized the way individuals, groups and communities communicate with each other. In January 2016, GlobalWebIndex[1] estimated the number of internet users to be around 3.42 billion, 2.31 among them are active Social Media users. They reached an annual growth of more than 10% between 2015 and 2016. Figure 1.1 presents the number of active users in the best-known Social Media websites as of April 2016[2]. The presented websites amounts hundreds of millions of users and some of them reached/surpassed 1 billion active users. Facebook takes the lead with over 1.59 billion active users. According to the same source, more than 97% of US citizens aged between 18 and 34 are registered to it. This unprecedented volume, penetration and variety of user-generated content constitute golden opportunities for understanding social behavior and building intelligent systems.

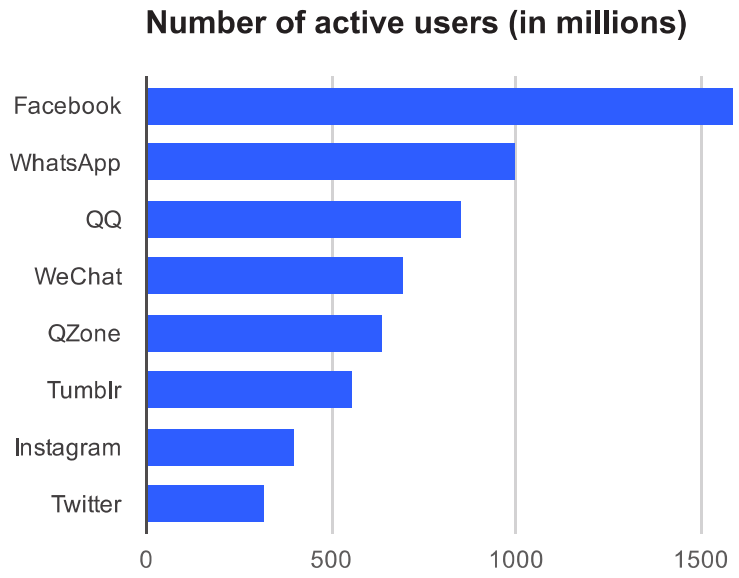


Figure 1.1: The number of active users in the best-known Social Media websites as of April 2016 (source: www.statista.com).

[1] www.globalwebindex.net
[2] www.statista.com

Social Media Mining is the field that aims to extract actionable knowledge from these Social Media data. It is receiving an increasing interest from various domains. Research in Social Media Mining uses techniques and algorithms from computer science to model and extract patterns related to social science. In fact, the field is deeply related to Semantic Web, Network Analysis, Machine Learning, Natural Language Processing, Statistics, Sociology, Ethnography, etc. Using the tools provided by all these domains, many research applications have emerged such as: community detection (Leskovec et al., 2010), spam identification (Hu et al., 2014), social recommendation (Tang et al., 2013), information diffusion (Guille et al., 2013), influence and lurking detection (Tagarelli and Interdonato, 2014), trust and distrust analysis (Tang et al., 2015), expertise prediction (McAuley and Leskovec, 2013), sentiment classification (Kiritchenko et al., 2014), etc.

Social Media Mining has plenty of use cases in business, politics, management, etc. It facilitates the elaboration of more targeted marketing campaigns, enables the tracking of the citizens opinions, performs predictive analysis to help managers in their decision making, etc. One of the challenges facing Social Media Mining is to extract useful contextual knowledge from the unstructured social data[3]. Indeed, Social Media Mining techniques must handle this unprecedented heterogeneity (variety), scale (volume), speed (velocity), trust (veracity), privacy and accuracy coming along with social data (Che et al., 2013). Despite theses challenges, most organizations want to capture the context of these unstructured data and put them side by side with their structured data for a clearer picture. The context of Social Media data may be studied through different dimensions. Some of these dimensions are presented in figure 1.2 as research questions and summarized below.

1.1.1 Who is talking?

Research filling under this question aims to mine characteristics of Social Media users. These characteristics define the user social role (Forestier et al., 2012), for example: an expert (Guy et al., 2013), an influencer (Agarwal et al., 2008), a lurker[4] (Tagarelli and Interdonato, 2014), a reputable user (Li et al., 2015), etc. The applied methods are usually based on graph theory using the structure of Social Media data (Zafarani et al., 2014). Knowing the user social role may be useful for a variety of purposes. First, online forum moderators are often interested in the identification of expert users to facilitate the access to the best answers that have more chances of being correct and informative. Then, companies usually target influencers in their marketing campaigns for a better diffusion of their offerings. After that, identifying lurkers is the first step in building strategies that would encourage them to de-lurk and become active in the community. Finally, the user reputation indicates the trustworthiness of his/her posted messages or selected ratings

[3] www.slideshare.net/simplify360/big-data-and-social-media-analytics

[4] A passive user who observes but does not actively participate in the community.

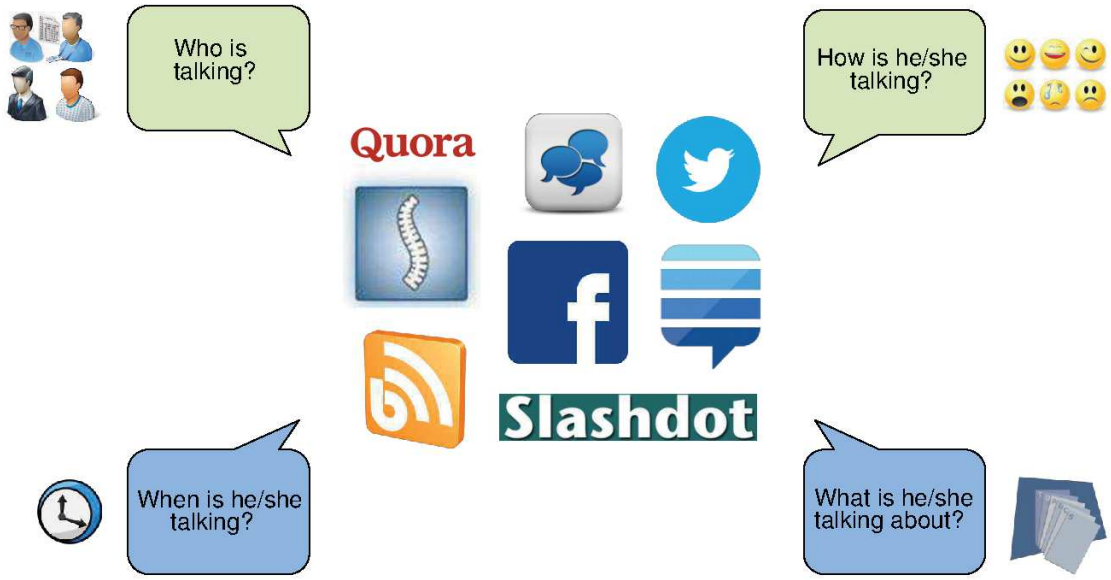


Figure 1.2: The dimensions of the contextual knowledge that can be extracted from social data.

1.1.2 How is he/she talking?

This dimension includes detecting the expressed opinions and affect states (Pang and Lee, 2008). Research in sentiment analysis includes: subjectivity detection (Riloff et al., 2005), polarity classification (Socher et al., 2013), emotion identification (Mohammad and Kiritchenko, 2015), intensity measurement (Kiritchenko et al., 2016), etc. The applied methods may be supervised (Pang et al., 2002) or unsupervised (Turney, 2002). On the one hand, supervised classification models may be trained on annotated documents in order to identify the expressed sentiment class (Mohammad et al., 2013). On the other hand, unsupervised sentiment classification approaches do not need annotated documents. They are usually based on the elaboration and/or use of sentiment lexicons (Hu et al., 2013). Sentiment analysis has many real-world applications. For example, brands and politicians are interested in detecting the opinions expressed in Social Media about their products or their programs. Therefore, the field is receiving much attention and effort from both scientific and economic communities.

1.1.3 What is he/she talking about?

Research answering this question aims to identify the topic of the discussion. Supervised and unsupervised classification methods are also applied for this purpose. First, supervised classification models may be trained on annotated documents in order to identify the discussed topic on new ones (Kinsella et al., 2011). In this case, already classified documents are usually used to learn these models. On the other

hand, unsupervised classification techniques do not need annotated data to group similar documents treating the same topics (Rosa et al., 2011). The work done to identify the subject of the discussion in Social Media includes: categorizing lay requests to web-based expert forums (Himmel et al., 2009), classifying questions in Question-Answering websites to the adequate category (Qu et al., 2012), grouping tweets according to the discussed topic (Yang et al., 2014), etc.

1.1.4 When is he/she talking?

It may be interesting to study the three above mentioned dimensions according to time in order to track the evolution of a given target, let trends emerge and detect breakpoints. For example, trust prediction and rating websites can study the evolution of individual trust over time (Tang et al., 2012). Similarly, recommender systems may consider the change of the user taste based on his/her online reviews (McAuley and Leskovec, 2013). Finally, topic models may include the evolution of the expressed sentiments about the considered topic (Dermouche et al., 2014), etc.

The techniques applied in each dimension depends on the nature of the studied social data. For example, social networks are represented by network structures to denote different relationships, while online forums are organized into threaded discussions to organize the posted textual contents. Network-based Social Media Mining uses graph theory in order to represent the network structure of social data and ignores the exchanged textual messages (Zafarani et al., 2014). However, the study of the textual contents may reveal complementary knowledge that can not be extracted using the network-based analysis (Farzindar and Inkpen, 2015). Furthermore, among the work that uses the textual contents, most of them concerned the English language. Therefore, this thesis describes some contributions in developing methods and resources for French Social Media Mining.

In this thesis, we propose contributions answering the first two questions : *« Who is talking? »* and *« How is he/she talking? »*. The textual content of the exchanged messages is used in developing approaches answering these questions. Our goal is to propose and evaluate methods and resources dealing with these two dimensions for the French language. Regarding the first dimension, the aim is to define the social role of Social Media users. In our case, we are interested in predicting the expertise (Abdaoui et al., 2014) and reputation (Abdaoui et al., 2015a) of online forums users. Indeed, the identification of the user expertise and reputation facilitates the access to the best answers that have more chances of being correct and informative. In more general terms, measuring the users notoriety in Social Media, can be an important differentiating factor between these users. Regarding the second dimension, the aim is to provide tools and resources in order to detect affect states in French Social Media contents. Therefore, we develop and evaluate methods (Abdaoui et al., 2015b) and resources (Abdaoui et al., 2016) for the classification of French text documents (tweets, product reviews, etc.) according to their polarity, subjectivity and emotion.

1.2 Research Contributions

In order to answer the above mentioned research questions (who and how), the following technical, methodological, theoretical and experimental contributions have been made.

1.2.1 Data Collection and Annotation

In order to carry out the empirical studies on real social data, a collection step has been done. On the one hand, a crawler has been implemented to collect online French forums. Four forums related to health issues have been collected. The implemented crawler parsed all the publicly available information about the posted messages (textual content, posting date, posting time, etc.), their authors (pseudonym, age, rank, etc.) and their threads (creation date, creation time, creating user, etc.). On the other hand, the Twitter API has been queried either to collect tweets containing some keywords or to retrieve specific tweets annotated in reference benchmarks. Other sentiment benchmarks have been downloaded directly from the internet.

Furthermore, manual annotations have been performed on the collected data in order to evaluate the proposed methods and the compiled resources. All the annotated items have been checked by multiple human annotators and agreement measures have been computed.

1.2.2 Methods for User Expertise and Reputation

First, we propose and evaluate a content-based method for the prediction of the user expertise in online forums. The method uses forums that hire medical experts as training data in order to learn supervised classification models. We conduct diverse experiments in order to evaluate if the models learned on a given website may be used efficiently on other websites. Two real French forums have been used in these evaluations.

Then, we suggest to estimate the reputation of these users based on the replies addressed to them. Therefore, we propose a rule based heuristic to find the recipient of each forum message. This heuristic extracts a multi-graph from each thread discussion, where the nodes represent the users and the edges represent the exchanged messages between these users. The proposed heuristic is based on specifically designed rules such as: the presence of pseudonyms inside the posted message, the presence of a question posted before the current message, etc. Using the extracted network, we propose a new score that counts the number of replies expressing trust and those expressing distrust. In order to give more importance to the replies posted by trusted users, we suggest to weight each reply by the reputation of its author. Therefore, the proposed reputation score can be computed by iterating until convergence. The rule based heuristic and the reputation score have been evaluated on two real French forums.

1.2.3 French Sentiment Lexicon

A new French sentiment lexicon has been compiled considering both the polarity and the emotion of French terms. It has been created following a new semi-automatic protocol for translating and expanding lexical resources in different languages. Indeed, we translated and expanded to synonyms the English NRC Word Emotion Association Lexicon (NRC-EmoLex). Online translators have been automatically queried in order to create a first version of our new French Expanded Emotion Lexicon (FEEL). Then, a human professional translator manually validated the automatically obtained entries and the associated emotions. Manual annotations have been performed to validate the chosen sentiment by multiple annotators. Diverse experiments have been conducted to compare the final version of FEEL with other existing French lexicons from the literature on reference benchmarks.

1.2.4 Sentiment Classification Process

Extensive experiments have been conducted on reference benchmarks used in previous French sentiment classification challenges. We define a feature engineering process to choose the best combination of pre-processings, features, resources and parameters for each benchmark. This process is applied using a 10-fold cross validation on the training set of each benchmark. The chosen configurations obtained comparable results to the best performing systems at each challenge. The conducted experiments have shown a clear contrast between the best performing features and methods for short and long documents. The implemented features and methods along with the learned classification have been made available online on a dedicated web-platform.

1.3 Thesis Organization

The organization of the remaining of this dissertation is presented in figure 1.3. The chapters are summarized as follows. First, chapter 2 presents the state of the art methods in Social Media Mining that fill into the raised research questions. First, it briefly summarize studies using the network structure of the social data. Then, it describes Text Mining steps, tasks and techniques in social data. Finally, Sentiment Analysis is studied as a use case of Text Mining in Social Media. Then, the chapters describing our contributions are organized in two parts. Each part correspond to a research question.

1.3.1 Part I: User Expertise and Reputation

In this first part, we are interested in mining the trustworthiness of users in online forums. It is composed of two chapters answering the question *« Who is talking? »*.

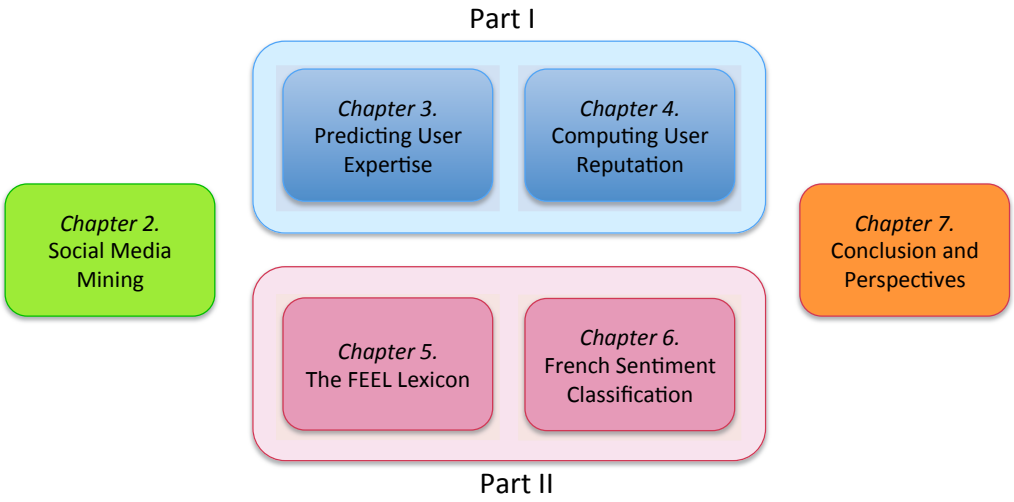


Figure 1.3: The organization of the remaining of this dissertation.

First, chapter 3 presents a content based method to predict the expertise of the author of a given forum post. The goal consists in categorizing posts written by medical experts from those written by laymen. The proposed approach uses forums that hire medical experts and indicate explicitly their role on the website in order to learn supervised classification models. These models will be able to predict the medical expertise in any other forum post. The textual content of the posts has been used in order to represent the posts and learn various classification models. After presenting the used French forums, we describe the proposed steps to perform the categorization (pre-processing, annotation and classification). Then, we present the experiments conducted by cross validation on each forum and those conducted by learning on one forum and testing on the other one. Finally, we discuss the obtained results and provide an error analysis step.

Chapter 4 describes a second contribution answering the same research question *« Who is talking? »*. The aim is to estimate the user reputation in online forums. In order to compute the reputation of a given user, we suggest to look at the posts that are addressed to him rather than his/her own posts. The idea is to detect expressions of agreement and thanking for positive replies and expressions of disagreement and depreciation for negative ones. In order to compute such reputation, we need to know the recipient(s) of each forum message. Unfortunately, the recipient is rarely known through the structure of the forum. Therefore, a rule based heuristic has been proposed and evaluated in order to construct a multi-graph where the nodes represent the users and the edges represent the replying messages between these users. Then, each reply has been evaluated in order to detect positive and negative replies. Finally, we propose a reputation measure inspired from the PageRank algorithm. This measure takes into account propagation aspects by giving more weight to replies posted by reputable users. This approach has been implemented and evaluated on two French forums.

1.3.2 Part II: Sentiment Analysis

This second part is composed of two chapters answering the question *« How is he/she talking? »*. The goal is to identify affect states in French text documents.

Chapter 5 concerns the compilation of a new French sentiment and emotion lexicon following the Ekman typology (Ekman, 1992). On the one hand, this chapter describes the semi-automatic compilation process of our new French Expanded Emotion Lexicon (FEEL). Indeed, this lexicon is obtained semi-automatically by translating and expanding to synonyms the well-known English lexicon NRC EmoLex (Mohammad and Turney, 2013). First, the english entries are translated and expanded to synonyms automatically by querying online translators. Then, the automatically obtained French entries are checked by a professional human translator. On the other hand, it presents extensive evaluations on French benchmarks for polarity and emotion classification in order to compare FEEL with existing French lexicons.

Chapter 6 concerns the evaluation of the state of the art features, methods and resources for French sentiment classification. Benchmarks released in previous French sentiment classification challenges have been used in these evaluations. The considered textual documents consist in tweets, product reviews and debate reports. These documents have been classified according to their polarity, subjectivity and emotion. A feature engineering process has been applied by cross validation in order to choose the best pre-processings, features and parameters for each benchmark and each classification task. Finally, this chapter discusses the best features and methods for French sentiment classification according to the text nature and the studied task.

Finally, chapter 7 concludes this thesis and discusses its various perspectives for future research directions.

1.4 Publications

The following published, in press or accepted papers are partial outputs of this thesis. Table 1.1 organizes them according to the studied thesis part.

1.4.1 International Journals (2)

1. Amine Abdaoui, Jérôme Azé, Sandra Bringay and Pascal Poncelet. FEEL: French Expanded Emotion Lexicon. Language Resources and Evaluation, pp. 1–23, 2016 (**Impact Factor: 0.975**, published online).

2. Amine Abdaoui, Jérôme Azé, Sandra Bringay and Pascal Poncelet. Expertise in French Health Forums. Health Informatics Journal, 2016 (**Impact factor: 1.578**, accepted).

Two other journal papers will be submitted soon.

1.4.2 International Conferences (6)

3. Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet. Collaborative Content-Based Method for Estimating User Reputation in Online Forums. Proceedings of the 16th International Conference on Web Information Systems Engineering, WISE 2015: (2) 292-299 [CORE: A].

4. Amine Abdaoui, Jérôme Azé, Sandra Bringay and Pascal Poncelet. E-Patient reputation in Health Forums. Proceedings of the 15th World Congress on Health and Biomedical Informatics, MEDINFO 2015: 137-141 [CORE: B].

5. Amine Abdaoui, Jérôme Azé, Sandra Bringay and Pascal Poncelet. Assisting e-patients in an Ask the Doctor Service. Proceedings of the 26th Medical Informatics Europe Conference, MIE 2015: 572-576.

6. Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar and Pascal Poncelet. Predicting Medical Roles in Online Health Fora. Proceedings of the second Statistical Speech and Language Processing Conference, SLSP 2014: 247-258.

7. Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar, Pascal Poncelet: Analysis of Forum Posts Written by Patients and Health Professionals. Proceedings of the 25th Medical Informatics Europe Conference, MIE Posters 2014: 1185.

8. Soumia Melzi, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet and Florence Galtier. Patient's rationale: Patient Knowledge retrieval from health forums. Proceedings of the the 6th International Conference on eHealth, Telemedicine, and Social Medicine, ETELEMED 2014: 140-145.

1.4.3 French Conferences (2)

9. Amine Abdaoui. Nouvelle méthode de calcul de la réputation dans les forums de santé. Actes des 16èmes Journées Francophones sur l'Extraction et Gestion des Connaissances, EGC 2016: 231-236.

10. Soumia Melzi, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Pascal Poncelet, Florence Galtier. Que ressentent les patients ? Actes des 14èmes Journées Francophones sur l'Extraction et Gestion des Connaissances, EGC 2014: 449-454.

1.4.4 Workshops (2)

11. Amine Abdaoui, Mike Donald Tapi Nzali, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi and Pascal Poncelet. ADVANSE : Analyse du sentiment, de l'opinion et de l'émotion sur des Tweets Français. Actes de la

22e conférence sur le Traitement Automatique des Langues Naturelles TALN 2015, 11ème Défi de Fouille de Textes, DEFT 2015: 78-87.

12. Amine Abdaoui, Jérôme Azé, Sandra Bringay, Natalia Grabar and Pascal Poncelet. Analyse des messages des patients et des médecins dans les fora de santé. 25es Journées Francophones de l'Ingénierie des Connaissances IC 2014, Atelier Intelligence Artificielle et Santé.

Table 1.1: A summary of the publications related to each thesis part.

	Part I: User Expertise and Reputation	**Part II: Sentiment Analysis**
International Journals	[2]	[1]
International Conferences	[3], [4], [5], [6], [7]	[8]
French Conferences	[9]	[10]
Workshops	[12]	[11]

Chapter

2 Social Media Mining

Contents

2.1 Introduction 14

2.2 Network Analysis in Social Media 14

2.2.1 Extracting Social Networks 15

2.2.2 Mining the Expertise in Social Networks 17

2.3 Text Mining in Social Media 18

2.3.1 Linguistic Pre-processing 19

2.3.2 Text Representation 20

2.3.3 Text Categorization 21

2.4 Sentiment Analysis 25

2.4.1 Lexicons 26

2.4.2 Classification Methods 30

2.4.3 Benchmarks 33

2.5 Conclusion 35

2.1 Introduction

Social Media Mining integrates social theories with computational methods to study how individuals interact and how communities form. It uses theoretical methodologies from various disciplines such as computer science, network analysis, sociology, statistics, mathematics, etc. It is the process of representing, analyzing and extracting meaningful patterns from Social Media data (Zafarani et al., 2014). Many research applications have emerged under the umbrella of Social Media Mining including: community detection (Leskovec et al., 2010), social recommendation (Tang et al., 2013), influence and lurking (Tagarelli and Interdonato, 2014), sentiment analysis (Pang and Lee, 2008), etc.

Social Media data is characterized by its enormous size, its unstructured form and its abundant social relations. These social relations are usually represented in network structures (followees-followers, friendship, etc.). Therefore, much research in Social Media Mining is based on graph theory as detailed in the following book (Zafarani et al., 2014). However, the textual content of the posted messages may reveal a lot of interesting patterns that could not be extracted using network structures (Farzindar and Inkpen, 2015). In this thesis, more attention will be given to the application of Text Mining techniques to investigate the user-generated textual contents. This chapter is dedicated to the state of the art of the methods related to the main contributions of this thesis.

The remainder of this chapter is organized as follows. First, a brief state of the art of network analysis in Social Media is presented focusing on the mining of user expertise. Then, a more detailed description of Text Mining in Social Media is introduced. We focus on text categorization since this task is used many times in our contributions. Finally, we present a literature review of research in sentiment analysis, which is an illustration of content-based Social Media mining.

2.2 Network Analysis in Social Media

Social data is usually represented in network structures (graphs, multi-graphs, etc.). As an example, let us consider a set of three individuals John, Lucie and Mike on a given social network. Each individual can be represented using a node. An arrow (a directed edge) is created from an individual to another if the first follows the later on the given social network. For example, let us consider that: (i) John follows Lucie; (ii) Lucie follows John and Mike; and finally (iii) Mike follows John. This network can be represented by an oriented graph as shown in figure 2.1.

Graphs representing social data can be directed or not according to the considered relations (follows, friendship, etc.). Furthermore, the nodes may have additional attributes (the name of the user, his/her age, etc.) and the edges may be weighted (counting the number of replies, etc.). Using these graphs, network based analysis can be applied (Brandes and Erlebach, 2005). For instance, we can say that John

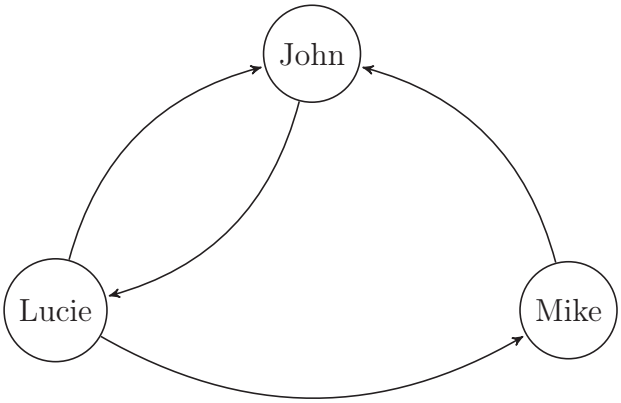


Figure 2.1: A simple social network of three users represented by an oriented graph.

is the most important individual since he is followed by two individuals while Lucie and Mike are followed by only one individual. The remainder of this section is divided into two parts. First, we describe a summary of the work done to extract these networks from various sources. Then, we summarize studies that used these networks in order to compute user expertise and reputation.

2.2.1 Extracting Social Networks

If most of the work done in Social Media Mining use graph theory, many social websites can not be directly represented using the above mentioned structures. Therefore, many studies concerned the extraction of a network from diverse sources (emails, scientific publications, online forums, etc.). Here, we describe studies that dealt with each of these sources. A particular attention is given to the extraction of interaction networks from online forums.

(Bekkerman and McCallum, 2004) proposed a statistical and learning based system that identifies people in email exchanges, finds their web presence and automatically fills the fields of a contact address book using Conditional Random Fields. The system builds a user's social network by recursively calling itself on new people discovered on the web. (Matsuo et al., 2007) proposed a web-based system that extracts social networks from scientific publications. Four relations have been extracted: co-authors, lab members, project members and conference participants. Therefore, multiple relations can exist between two nodes (the extracted network is a multi-graph). The system computes co-occurrences of names in order to extract these relations. Online discussion communities such as forums have also been studied in order to extract interaction networks. The studied relation is usually 'answer-to' or 'reply-to'. Most of methods found in the literature use the HTML structure of the web page in order to identify explicit message quoting (Welser et al., 2007, Zhang et al., 2007, Adamic et al., 2008, Stavrianou et al., 2009). However, explicit quoting functionality is not always provided in online forums, and even when

it exists many discussion participants do not use it. Moreover, a message may have many recipients. Consequently, posting it as an answer to another specific message may be insufficient.

While the HTML quoting structures are rarely provided and poorly used in online forums, the posts textual contents may reveal a lot of information regarding their recipient(s). (Gruzd and Haythornthwaite, 2008) presented an automatic approach to discover and analyze social networks from threaded discussions in online courses. The authors proposed a Name Entity Recognition system to extract name mentions inside the textual content of posts. After a pre-processing step (removing quotations, stop words, etc.), their method used a dictionary of names combined with manually designed linguistic rules. For example to recognize nicknames, abbreviations and misspellings, they relied on the context words (Hi, dear, etc.). To exclude names of buildings, organizations, etc. they ignored sequences of more than two capital words. To remove street names and avenues, they checked if names are followed by a prohibited list of words (Street, Ave, etc.).

Another textual based method to extract a network of user interactions from online forums has been proposed in (Forestier et al., 2011). They suggested to infer three types of interactions: structural relations, name citations and text quotations. While structural relations can be inferred directly from the structure of the forum, name citations and text quotations require analyzing the textual contents. On the one hand, name citation relations have been extracted by searching pseudonyms of authors inside the posts. The Levenshtein distance normalized by the pseudonym length (Soukoreff and MacKenzie, 2001) has been used to take into account misspelled or incomplete pseudonyms. Moreover, the authors suggested that users often use non-existent words as pseudonyms. Therefore, they added the constraint that searched pseudonyms must be unknown to the TreeTagger dictionary (Schmid, 1994). On the other hand, text quotations are extracted by comparing sequences of words inside a message and the messages that have been posted before in the same thread (whether these sequences appear between quotation marks or not). In this case, the number of words must be chosen to maximize both the precision and the recall.

If all these methods are relevant to extract social networks from online discussions, we have developed new rules and implemented them into a smart heuristic to capture new types of interactions such as grouped posts dedicated to all the persons participating in the thread, questions and answers, etc. This heuristic will be detailed in in section 4.2.3 page 56. Once the social data is represented using graph structures, network-based analysis techniques can be applied. In our case, we are interested in methods allowing the identification of expert users inside social networks, which corresponds to the first research question raised in the case of this thesis (Who?).

2.2.2 Mining the Expertise in Social Networks

An expert is a user who has knowledge about the discussed topic and as such his/her posted information can be trusted (Forestier et al., 2012). Many studies on social network analysis took advantage of the work done in ranking authoritative web pages. The best-known algorithm is PageRank (Page et al., 1999), which ranks web pages according to their importance. The basic idea behind this algorithm is to give more importance to web pages that are pointed by many other pages (especially authoritative ones). The rank of a web page is computed as the mean of the ranks of pointing web pages divided by the out-degree[1] of each pointing page. For example, using this algorithm the authority of the user John in figure 2.1 will be computed as follows:

$$Authority(John) = \frac{1}{2}\left(\frac{Authority(Lucie)}{2} + Authority(Mike)\right)$$

The equation is recursive and may be computed by starting with any set of values and iterating until it converges. Few years later, (Gyöngyi et al., 2004) proposed TrustRank in order to separate good pages from web spams using a semi-automatic method. First, they selected a small set of seed pages to be evaluated by an expert. Then, they used the PageRank algorithm to propagate the trust and discover other pages that are likely to be trusted.

These two algorithms have been used in identifying expert users in social networks. For example, (Matsuo et al., 2004) proposed two values of trust for each scientific researcher: (i) a social trust which uses a PageRank like model to compute the authority of a researcher on the community; and (ii) an individual trust which uses TrustRank by starting with a source researcher and propagating his/her trust. (Heidemann et al., 2010) used a PageRank model in order to find key users on a publicly available Facebook dataset. These users are often targeted by companies in the case of advertising strategies.

Regarding the identification of expert users in online forums, (Zhang et al., 2007) evaluated a set of ranking algorithms on a network extracted from a Java forum. The extracted network is represented by a directed graph where the nodes represent the users and the edges represent the 'reply-to' relations between these users. They proposed an adaptation of PageRank called ExpertiseRank. The basic idea is to compute how many people a user helps and how much expertise these people have. The intuition is that if B is able to answer A's question, and C is able to answer B's question, then C's expertise should be boosted because he was able to answer a question of someone who himself/herself has some expertise. They compared different techniques with a gold set of users ranked manually and obtained the best results using ExpertiseRank. However, this method do not take into account the content of the answer itself nor its correctness.

[1]Number of outbound links

Similar approaches to discover experts in online discussions using network-based analysis have been proposed in (Widén-Wulff et al., 2008,Kardan et al., 2011,Kardan et al., 2012). These studies ignores the textual content of the posted messages. However, users may send many irrelevant or empty messages which may cause some mistakes in finding experts. Therefore, the analysis of the posts content may be used (Wanas et al., 2008) or combined with link analysis (Omidvar et al., 2014, Rafiei and Kardan, 2015) in order to infer the user expertise. Text Mining and Natural Language Processing techniques can be applied in order to perform this task. In the next section, we present a detailed description regarding the application of Text Mining techniques on data from Social Media. We focus on text categorization which allows the classification of text documents in pre-established classes.

2.3 Text Mining in Social Media

Text Mining is the computational process of discovering new patterns (or knowledge) from unstructured text documents (Stavrianou et al., 2007). As shown in figure 2.2, the text mining process is similar to the general data mining one. First, textual data is collected from various sources including Social Media, news articles, biological literature, etc. Then, an important effort should be deployed in cleaning and pre-processing the collected data by applying Natural Language Processing techniques. After that, Machine Learning and Data Analysis methods are applied in order to learn statistical models. Finally, the learned models are evaluated and interpreted in order to extract useful and unknown knowledge. These steps will be detailed in the following.

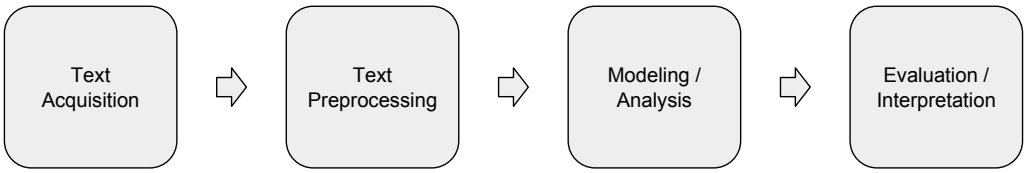


Figure 2.2: The Text Mining process.

Hot topics in Text Mining include Language Modeling (Le et al., 2011), Machine Translation (Son et al., 2012), Text Summarization (Pal and Saha, 2014), Text Classification (Lamirel et al., 2015), etc. In this thesis, we focus on Text Classification or Text Categorization which will be used many times in our methodologies. It consists in assigning textual documents to predefined classes. However, before applying classification techniques, many language issues should be considered in order to clean, transform and represent the unstructured textual data. The remainder of this section is divided as follows. First, we present the linguistic pre-processing steps that are usually applied on Social Media texts. Then, we describe some possible representations of text documents. Finally, we detail the task of Text Categorization (the used models, evaluation metrics and annotation steps).

2.3.1 Linguistic Pre-processing

Natural Language Processing (NLP) is the field that allows computers to understand and process human (natural) language. Indeed, the unstructured and complex nature of textual data makes the direct application of Machine Learning methods impossible. Furthermore, texts from Social Media have several linguistic peculiarities that may influence the classification performance (Farzindar and Inkpen, 2015). They contain many misspellings, abbreviations, repeated characters (enooooooough), repeated punctuation signs (!!!!!!!!), unconventional capitalizations (TIRED), slang words (lol), emoticons (:-)), etc. Therefore, many NLP based pre-processings can be applied to deal with these issues (Balahur, 2013). Some of them are summarized in the following.

- **Spelling Correction:** Correcting spelling mistakes may be performed using dedicated tools (such as Aspell[2]). These tools use a dictionary or a thesaurus for a given language. If a word has not been found in the dictionary it is considered as incorrect and replaced by its closest word in the dictionary (hapy → happy);
- **Normalization:** Hyperlinks, emails or user pseudonyms are usually normalized and replaced by a specific term (abdaoui@lirmm.fr → email). Indeed, it is more interesting to know that an email is mentioned rather than whose email is mentioned. This normalization can be performed with specifically designed regular expressions. Its application depends on the text nature and on the desired task;
- **Stop Words:** Common words that are frequently used in the text (to, of, etc.) may be filtered out according the text mining task. Plenty of stop lists for various languages can be found on the internet. One can also customize his own stop words list according to the words that may or may not be interesting for the desired task;
- **Slang Replacement:** In recent years, some slang words have been widely used in Social Media. These slang words affect the semantic analysis of natural language, because most existing resources are designed for well-written language (O'Connor et al., 2010). A simple way to replace slang words with their corresponding text expressions (lol → lot of laugh) is to use pre-established lists;
- **Stemming and Lemmatization:** Stemming is the process of reducing inflected forms of a word to their root (stem). For example the words 'democratic' and 'democratization' have the same stem which is 'democrat'. Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words. It is the process of replacing the inflected forms

[2]www.aspell.net

by their canonic form (the infinitive for verbs and the third person singular for the remaining part of speech components). For example: the words 'is' and 'are' have the same lemme which is 'be';

- **Part Of Speech Tagging:** Part Of Speech (POS) tagging is the process of marking up a word in a text as corresponding to a particular category. For example: noun, verb, adjective, adverb, preposition, etc. Current POS tools implement tagging algorithms which may be rule-based or stochastic;

- **Segmentation and Tokenization:** Segmentation is the process of dividing a text document into meaningful units, such as sentences or phrases. Most segmentation tools use punctuation signs (mainly the full stop) with some heuristics to avoid non trivial cases such as 'Mr. Smith'. Tokenization usually refers to the segmentation of text into small and independent tokens (words, numbers, etc.). A list of delimiter (' ', '.', etc.) can be customized according to the desired output. Tokenization is essential in order to transform the unstructured text documents into a form that can be processed by Machine Learning techniques.

These linguistic pre-processings are known to have a big influence on the Text Mining results as it will be demonstrated in this thesis. Once these pre-processings are applied, the unstructured textual data should be represented in a computational way as it will be described in the next section.

2.3.2 Text Representation

In order to apply Machine Learning methods, text documents should be converted into "processable units". Indeed, we need an efficient document representation in order to build effective classification models. One of the basic representation methods for text documents is the Bag Of Words (BOW) representation (Salton et al., 1975). This method represents each document by a vector of fixed size using the frequency count of its words. It is also called the Vector Space Model (VSM).

Let $D = d_1, d_2, ..., d_n$ be the set of textual documents in the data. Let $F = f_1, f_2, ..., f_m$ be the set of vocabulary words (features) used in the data. Then, VSM represents each document d_j by the vector $(w_{1j}, w_{2j}, ..., w_{mj})$, where w_{ij} is the weight of the feature f_i in the document d_j. Many weighting schemes can be considered such as:

- **Boolean:** w_{ij} takes 1 if the feature f_i appears in the document d_j, 0 otherwise;

- **Frequency:** w_{ij} takes the frequency of the feature f_i in the document d_j;

- **TF-IDF:** w_{ij} takes the frequency of the feature f_i in the document d_j multiplied by the inverted frequency of the documents where f_i appears.

Extensions of the BOW model have been proposed in order to take into account multiple tokens (ngrams) rather than words which captures some additional information about the order and the position of the words. However, this model has other limitations mainly related to the high dimensionality of the representation and the lack of generalization.

To reduce the high dimensionality of the representation, feature subset selection algorithms can be applied in order to filter out inadequate and redundant features. In addition to a significant gain in computational time, feature selection often improves the classification results by removing noisy features. Feature selection methods can be organized into two main categories:

- **Filter methods:** evaluate the intrinsic properties of features, without applying any classifier (Guyon and Elisseeff, 2003). Usually, they evaluate the worth of the features on the base of their statistical properties (for example, by measuring the information gain with respect to the class). These methods are particularly effective in computation time and robust to over-fitting;

- **Wrapper methods:** learn classifiers on many subsets of the features and select the subset that induces the best results (Kohavi and John, 1997). Since it is very costly to test all the feature subsets, these methods rely on heuristic search of the space of these subsets.

To overcome the lack of generalization, recent approaches suggest to use a contiguous word representation that captures much more semantic information (Collobert et al., 2011). In this representation, each word is represented by a vector of a fixed size (for example 100 or 500 dimensions). These vectors, also called Word Embeddings, are learned using Deep Neural Networks on large untagged datasets based on the syntactic context of words (Mikolov et al., 2013b). Since 'coffee' and 'tea' appear in the same contexts, they should obtain very close vectors.

These representations are necessary before applying Text Mining techniques regardless of the text nature. The following section presents a basic task in Text Mining which is Text Categorization.

2.3.3 Text Categorization

Text Categorization consists in assigning textual documents to predefined classes. It can be applied to many tasks such as: documents organization (Larkey, 1999), spam filtering (Cormack, 2007), opinion classification (Mohammad et al., 2013), author profiling (Weren et al., 2014), etc. In this thesis we are interested in supervised Text Categorization, which will be used to classify the user expertise and sentiment class from French text documents. Supervised Text Categorization needs annotated documents (documents for which the classes are already known) in order to learn classification models. Then, these models will be able to classify new documents into the appropriate class.

Classification Models

Classical Machine Learning algorithms can be used in order to perform Text Categorization. Among the best-known algorithms, we can cite:

- **Support Vector Machines:** Which represent the documents as points in the feature space (Cortes and Vapnik, 1995) and seek to construct a hyperplane that can separate the classes. The best hyperplane is the one which maximizes the margins (the distance between the hyperplane and the nearest training-data points). In addition to linear classification, SVM can efficiently perform non-linear classification using kernels (mapping the inputs into high-dimensional feature spaces). Using SVM requires to set some parameters such as: the kernel function, the complexity parameter, etc.;

- **Naive Bayes:** A family of probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features (John and Langley, 1995). They combine the probability model with a decision rule. One common rule is to pick the hypothesis that is most probable; which is known as the maximum a posteriori decision rule;

- **Decision Trees:** These classification models aim to build a classification tree based on the learning data. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items. Different algorithms use different metrics for choosing this variable (Quinlan, 1993, Rokach and Maimon, 2005);

- **Neural Networks:** Artificial Neural Networks are inspired from biological nervous systems, such as the brain. They try to build a network composed of an input layer (data), an output layer (classes) and one or more hidden layers (Haykin and Network, 2004). The back-propagation algorithm is often used to train their weights. With the recent increase in computational power, there has been a big interest in building Deep Neural Networks having more than one hidden layer (LeCun et al., 2015).

Evaluation Metrics

In order to evaluate the learned classification models, the annotated dataset should be separated into two sets (at least): (i) a training set which is used to learn the classification models and (ii) a testing set which is used to evaluate the learned models on new unseen documents. One of the most known techniques to separate the dataset and evaluate the models is to perform k-fold cross validation. This validation technique randomly partitions the dataset into k equal size subsets. A single subset is used for testing, while the remaining k-1 are used as a training set.

This process is repeated k times so that each of the k subsets is used as a testing set exactly once.

Among the best known evaluation metrics, we present the precision (P), the recall (R) and the F1-measure (F1). These metrics can be computed for each class 'c' using the following formulas:

$$P_c = \frac{TP_c}{TP_c + FP_c} \qquad R_c = \frac{TP_c}{TP_c + FN_c} \qquad F1_c = \frac{2 \times P_c \times R_c}{P_c + R_c}$$

Where: TP_c is the number true positives for class 'c'; FP_c is the number false positives for class 'c' and FN_c is the number false negatives for class 'c'.

Once computed for each class, these measures can be either macro averaged or micro averaged (Tsoumakas et al., 2009). On the one hand, macro averaging gives equal weight to each class. It is computed as the arithmetic mean of every class results. On the other hand, micro averaging is used to deal with unbalanced datasets. It is computed using the corresponding precision, recall or F1-measure formulas on the sum of individual true positives, false positives and false negatives. Finally, another way to deal with unbalanced datasets is to compute a weighted average measure (also known as the label-frequency-based micro-averaging). It is calculated by weighting each class results by its proportion of documents in the test set. Note that the weighted average F1-measure is not necessarily equal to the harmonic mean of the weighted average precision and the weighted average recall.

The following equations present the formulas of the macro, micro and weighted average precisions, recalls and F1 measures:

$$P_{ma} = \frac{1}{n} \times \sum_{c=c_1}^{c_n} P_c \qquad P_{mi} = \frac{\sum_{c=c_1}^{c_n} TP_c}{\sum_{c=c_1}^{c_n} (TP_c + FP_c)} \qquad P_{wa} = \frac{1}{n \times d} \sum_{c=c_1}^{c_n} P_c \times d_c$$

$$R_{ma} = \frac{1}{n} \times \sum_{c=c_1}^{c_n} R_c \qquad R_{mi} = \frac{\sum_{c=c_1}^{c_n} TP_c}{\sum_{c=c_1}^{c_n} (TP_c + FN_c)} \qquad R_{wa} = \frac{1}{n \times d} \sum_{c=c_1}^{c_n} R_c \times d_c$$

$$F1_{ma} = \frac{1}{n} \times \sum_{c=c_1}^{c_n} F1_c \qquad F1_{mi} = \frac{2 \times P_{mi} \times R_{mi}}{P_{mi} + R_{mi}} \qquad F1_{wa} = \frac{1}{n \times d} \sum_{c=c_1}^{c_n} F1_c \times d_c$$

Where n is the number of classes in the dataset; d_c is the number of documents in class c and d is the total number of documents.

Manual Annotation

Text Categorization requires tagged text documents in order to learn and evaluate classification models. When these documents are not tagged, manual annotation can be performed by human annotators. The annotators will have to choose the appropriate category for each text document. Usually, each document is annotated by more than one person. Then, agreement measures can be calculated to evaluate the quality of the annotations. The more agreement the annotators have, the better are the annotations. One of the most used agreement measures is the Kappa coefficient. It is generally thought to be more robust than simple percent agreement calculation, since it takes into account the agreement occurring by chance. Cohen's kappa was first introduced in (Cohen, 1968) to compute the nominal scale agreement between two ratters. Its formula is the following:

$$Kappa = \frac{P_0 - P_e}{1 - P_e}$$

Where P_0 is the observed agreement between the two annotators and P_e is the probability of chance agreement. Their formulas are presented as follows:

$$P_0 = \frac{1}{n} \sum_{c=c_1}^{c_k} d(c) \qquad\qquad P_e = \frac{1}{n^2} \sum_{c=c_1}^{c_k} d_1(c) \times d_2(c)$$

Where:
n is the number of documents;
k is the number of categories;
$d(c)$ is the number of documents commonly assigned to the category c;
$d_1(c)$ is the number of documents assigned to the category c by the first annotator;
$d_2(c)$ is the number of documents assigned to the category c by the second annotator.

Then, (Fleiss, 1971) generalized this coefficient to measure the agreement among many annotators (more than two annotators). The kappa measure of agreement (Cohen or Fleiss) is equal to 0 when the amount of agreement is what would be observed by chance. It is equal to 1 when there is perfect agreement. The kappa measure may be negative which denotes that the raters have more disagreement than what would be expected to be observed by chance. However, it can not be superior to 1 (perfect agreement). For intermediate values, (Landis and Koch, 1977) suggested the following interpretations:

As mentioned before, Text Categorization have many applications. In particular, sentiment classification allows the identification of opinions and emotions expressed in textual documents. This task answers to the second research question studied in the case of this thesis (*How?*). In the next section, we will detail the used approaches, resources and benchmarks for sentiment classification.

Table 2.1: The interpretations of Kapa values.

Kappa	Interpretation
< 0	Poor
0 - 0.2	Slight
0.2 - 0.4	Fair
0.4 - 0.6	Moderate
0.6 - 0.8	Substantial
0.8 - 1	Almost perfect

2.4 Sentiment Analysis

Sentiment Analysis is a Text Mining field that allows the semantic evaluation of pieces of text according to the expressed sentiments and opinions. Due to its large number of applications, Sentiment Analysis has been applied to a variety of domains: politics (Anjaria and Guddeti, 2014), education (Klebanov et al., 2013), health (Melzi et al., 2014), etc. and on text documents from different nature: tweets (Jiang et al., 2011), news headlines (Rao et al., 2014), emails (Pestian et al., 2012), etc.

Most commonly, the term "Sentiment Analysis" is used to refer to the task of automatically classifying text units according to their polarity. However, it actually covers a larger number of tasks dealing with the detection of the general attitude of the text author towards a particular target (Liu, 2012). Indeed, the author attitude may be observed through its:

- **Polarity:** Positive, negative or neutral (Pang et al., 2002);
- **Subjectivity:** Objective or subjective (Riloff et al., 2005);
- **Emotions:** Joy, surprise, anger, fear, etc.(Mohammad and Kiritchenko, 2015);
- **Intensity:** Either discrete (Pang and Lee, 2005) or real-valued sentiment scores (Kiritchenko et al., 2016).

Whereas polarity and subjectivity may be studied using two or three classes, plenty of emotions can be considered. Psychologists have proposed a number of theories that classify human emotions into taxonomies. Some emotions are considered basic, whereas others are considered complex. A number of typologies have been proposed for basic emotions (James, 1884, Plutchik, 1980, Ekman, 1992, Parrott, 2001). The most used typologies are Ekman (6 basic emotion classes) and Plutchnik (8 basic emotion classes).

The presented author attitude (polarity, subjectivity, emotion, etc.) may be investigated at different granularities levels:

- **Document level:** The majority of sentiment classification methods deal with the whole text document (product review, forum post, etc.) (Turney, 2002);
- **Sentence level:** These methods deal with the sentences or clauses and determine the expressed attitude (Wilson et al., 2005);
- **Aspect level:** Aspect level methods apply fine-grained analysis. First, the aspects of the target should be extracted. Then, the sentiment towards these aspects are identified. For example, the sentence "The iPhone's call quality is good, but its battery life is short" expresses a positive opinion towards the quality call but a negative one towards the battery life (Pontiki et al., 2014).

Research in sentiment analysis includes many other tasks such as: the extraction of the opinion holder (Kim and Hovy, 2006), the identification of the target of the sentiment (Bringay et al., 2014), etc. In this thesis, we restrict our Sentiment Analysis study to the classification of texts from Social Media according to the expressed sentiment (polarity, subjectivity and emotion) at the document level. A particular attention is given to the development of methods and resources for French sentiment classification. Since the used textual documents are extracted from Social Media, the proposed methods should consider their specificities (see section 2.3.1).

The remainder of this section is organized as follows. First, a literature of English and French sentiment lexicons is listed. Then, we describe in sentiment classification methods which are usually based on Statistics and Machine Learning. Finally, we discuss the main compilation methods of sentiment benchmarks and present those available for the French language.

2.4.1 Lexicons

Sentiments are mainly conveyed by words, therefore, many studies tried to compile sentiment resources which organize lists of words, phrases or idioms into predefined classes (polarities, emotions, etc.). These resources (also known as sentiment lexicons) can be constructed using three main approaches. First, they can be compiled manually by assigning the correct polarity or emotion conveyed by each word. Crowd-sourcing tools and serious gaming are often used to get a large number of human annotations (Lafourcade et al., 2015a, Mohammad and Turney, 2013). Second, they can be compiled automatically using dictionaries. This approach uses a small set of seed terms for which the conveyed sentiments are known. Then, it grows the seed set by searching synonyms and antonyms using dictionaries (Strapparava et al., 2004). Finally, the third approach constructs sentiment lexicons automatically using corpora in two possible ways. On the one hand, it can use annotated corpora of text documents and extract words that are frequent in a specific sentiment class and not in the other classes (Kiritchenko et al., 2014). On the other hand, it can use non-annotated corpora along with a small seed list of sentiment words in order to discover new ones by computing collocations (Harb et al., 2008) or using specifically

designed rules (Neviarouskaya et al., 2011). Each type of these approaches has its own limitations. The manual approach is labour intensive and time consuming while the automatic ones are error prone.

Most sentiment resources have been compiled for English terms and only few lexicons have been designed for French. In the following, we present some English and French sentiment lexicons used in the literature. These lexicons are summarized in table 2.2.

English Lexicons

The following lexicons have been compiled for English terms:

- **General Inquirer:** Contains more than 11,000 English words labeled manually by 182 categories including polarity and some emotions (Stone et al., 1968);
- **WordNet Affect:** Contains only hundreds of English words labeled with their expressed polarity and emotion. It was created by manually identifying seeds (words whose associations with sentiments are known) and spreading these emotions to all their synonyms using WordNet (Strapparava et al., 2004);
- **MPQA:** Contains 8,222 English subjectivity words draws from the General Inquirer and other resources. Three polarities are considered: positive, negative and neutral (Wilson et al., 2005);
- **Bing Liu's Opinion Lexicon:** Contains around 6,800 English opinion words associated with their polarities (positive and negative). It was created automatically using a corpus-based approach (Qiu et al., 2009);
- **SentiWordNet:** Contains more than 117,000 English WordNet synsets in its 3.0 version. Each synset has been associated with a positivity score and a negativity score. It was created semi-automatically by combining the results produced by eight ternary classifiers learned on a small set of manually labelled terms (Baccianella et al., 2010);
- **NRC-EmoLex:** Contains more than 14,000 English terms labeled by the expressed polarity (positive or negative) and emotion (joy, trust, anticipation, sadness, surprise, disgust, fear or anger). The authors used Amazon Mechanical Turk[3] in order to obtain a large number of manual annotations in order to compile their resource (Mohammad and Turney, 2013);
- **NRC Hashtag Emotion Lexicon:** Contains real valued English words between 0 (not associated) to infinity (maximally associated) for each sentiment polarity and emotion class. It gathers 16,862 unigrams (words) that have been

[3]www.mturk.com

created automatically using a corpus based approach. The corpus has been obtained from Twitter by extracting tweets that contains the following hashtags: #joy, #sadness, #surprise, #disgust, #fear and #anger (Mohammad and Kiritchenko, 2015);

- **LIWC:** Contains about 4,500 English words labeled by many categories including polarity and emotion. It was created by combining other existing resources and by validating the categories manually by human judges (Pennebaker et al., 2015).

All these English resources consider the sentiment polarity but only five among them offer the exact emotional category. The most extensive English emotion lexicons are NRC-EmoLex (Mohammad and Turney, 2013) and the NRC Hashtag Emotion lexicon (Mohammad and Kiritchenko, 2015). These lexicons have proven their performance in several sentiment and emotion classification tasks (Kiritchenko et al., 2014, Mohammad, 2012, Rosenthal et al., 2015). Indeed, NRC-EmoLex has been built on the General Inquirer (Stone et al., 1968) and the WordNet Affect (Strapparava et al., 2004) lexicons. Concretely, NRC-EmoLex corrects their terms and add new unigrams and bigrams using the wisdom of the crowds. Furthermore, using this lexicon the authors obtained remarkable results in the evaluation campaigns SEM-EVAL 2013 (Nakov et al., 2013) and SEM-EVAL 2014 (Rosenthal et al., 2014). For all these reasons, this lexicon received a particular attention in this thesis.

French Lexicons

The following lexicons have been compiled for French:

- **Affects:** Consists of about 1,300 French terms described by their polarity (positive and negative) and over 45 hierarchical emotional categories. It was automatically compiled and includes other information such as the intensity and the language level (common, literary) (Augustyn et al., 2006);

- **CASOAR:** Contains polarized subjective terms in French. It consists of 270 verbs, 632 adjectives, 296 names, 594 adverbs and 51,178 expressions. It was manually constructed from several corpora (press articles, web comments, etc.). However, this resource is not publicly available (Asher et al., 2008);

- **Polarimots:** Contains 7,483 French nouns, verbs, adjectives and adverbs whose polarity (positive, negative or neutral) has been semi-automatically annotated. 3,247 words have been added manually and 4,236 words has been created automatically by propagating the polarities (Gala and Brun, 2012);

- **Diko:** Based on an online game with a purpose where players are asked to indicate the polarity and the emotion of the displayed expression. They can choose between three polarities (positive, negative and neutral) and 21 emotions.

They can also enter a new emotion term when the exact emotion meaning of the displayed expression is not present between the 21 choices. Therefore, this lexicon associates 555,441 annotated expressions to almost 1,200 emotion terms (Lafourcade et al., 2015a).

Table 2.2: A summary of English and French Sentiment Lexicons.

Lexicon	**Language**	**Number of entries**	**Polarity**	**Emotion**	**Reference**
General Inquirer	English	11,788 terms	Yes	Yes	(Stone et al., 1968)
WordNet Affect	English	606 terms	Yes	Yes	(Strapparava et al., 2004)
MPQA	English	8,222 terms	Yes	No	(Wilson et al., 2005)
Liu's Lexicon	English	6,800 terms	Yes	No	(Qiu et al., 2009)
Senti-WordNet	English	117,659 expressions	Yes	No	(Baccianella et al., 2010)
NRC EmoLex	English	14,182 terms	Yes	Yes	(Mohammad and Turney, 2013)
NRC Hashtag	English	16,862 terms	Yes	Yes	(Mohammad and Kiritchenko, 2015)
LIWC	English	4,500 terms	Yes	Yes	(Pennebaker et al., 2015)
Affects	French	1,792 terms and 51,178 expressions	Yes	Yes	(Augustyn et al., 2006)
CASOAR	French	1,348 terms	Yes	No	(Asher et al., 2008)
Polarimots	French	7,483 terms	Yes	No	(Gala and Brun, 2012)
Diko	French	555,441 expressions	Yes	Yes	(Lafourcade et al., 2015a)

Few French resources have been proposed, especially those dealing with emotions. If all of the French lexicons offer the sentiment polarity, only two consider the exact emotional category. The Affects lexicon (Augustyn et al., 2006) which

contains only 1,200 terms associated with more than 45 hierarchical emotions and Diko (Lafourcade et al., 2015a) which contains about 450,000 non-lemmatized expressions associated with almost 1,200 emotion terms (many synonyms exist). None of these lexicons follows a well-known emotional typology such as : the six basic emotion classes proposed in (Ekman, 1992) or the eight emotions wheel proposed by (Plutchik, 1980), etc. The two remaining lexicons CASOAR (Asher et al., 2008) and Polarimots (Gala and Brun, 2012) consider only the polarity and not the emotion. These four lexicons are the only resources dedicated to French that have been found in the state of the art. Furthermore, one of them (CASOAR) is not freely distrusted. All these observations highlight the lack of adapted French sentiment and especially emotion resources following a well-known typology.

2.4.2 Classification Methods

Sentiment Classification methods are often based on techniques from Statistics, Natural Language Processing and Machine Learning. They can be grouped into two main categories: Unsupervised and Supervised. In this thesis, more attention is given to supervised sentiment classification approaches.

Unsupervised Sentiment Classification

Unsupervised or lexicon-based approaches decide the sentiment of a text based on sentiment lexicons. As presented in the previous section, sentiment lexicons can be compiled by computing collocations. The popular method proposed in (Turney, 2002) illustrates well these kind of approaches. The authors proposed an algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). They computed the semantic orientation of adjectives and adverbs as the mutual information between the given term and the word "excellent" minus the mutual information between the given term and the word "poor". The calculation of the mutual information has been performed by computing collocations returned by the AltaVista search engine. Afterwards, each review is recommended or not according to the average semantic orientation of its terms.

A similar approach has been proposed in (Dray et al., 2009) to classify text documents from a specific domain. First, the authors queried the Google search engine with the name of the domain and two sets of positive and negative seed words (seven positive words and seven negative words). This operation allowed them to build 14 learning corpora (seven positive and seven negative corpora). Then, these corpora have been used to extract new potential positive and negative adjectives using association rules. After that, they used the AcroDef mutual information measure to filter out adjectives that are not correlated with the seed words. Finally, each document has been classified to the majority class by computing the difference between the number of its positive and negative adjectives.

Instead of compiling new lexicons, (Taboada et al., 2011) presented the Semantic Orientation CALculator (SO-CAL) which uses existing dictionaries and incorporates intensification and negation. The consistency of the used dictionaries has been checked by manual annotation using the Amazon Mechanical Turk. The scores of words appearing under the score of a negation term has been inverted. They has been either increased or decreased if the words appear under the scope of a modifier (very, slightly, etc.). The obtained results have shown that SO-CAL's performance is consistent across domains and on completely unseen data.

Many other studies concerned sentiment classification using unsupervised techniques (Paltoglou and Thelwall, 2012, Hu et al., 2013). All of them are based on the elaboration and/or use of sentiment lexicons. Their main advantage is that they do not need annotated documents for training. However, it has been demonstrated that supervised techniques which learn classification models from annotated datasets outperform unsupervised ones (Pang et al., 2002, Nakov et al., 2013). Furthermore, sentiment lexicons can be used as features in supervised sentiment classification (Mohammad et al., 2013, Rastogi et al., 2014, Hamdan et al., 2015).

Supervised Sentiment Classification

Most current sentiment classification techniques use supervised learning methods. The first work that considered this approach in classifying movie reviews according to their sentiment classes (positive and negative) has been proposed in (Pang et al., 2002). The authors showed that standard machine learning techniques definitively outperform human-produced baselines. Three machine learning methods have been employed: Naive Bayes, Maximum Entropy classification and Support Vector Machines (SVM). As features, they considered unigrams, bigrams and Part Of Speech tags. They have found that the classifiers performed better when a binary feature was used indicating the presence of a unigram in the text, instead of a numerical feature indicating the number of appearances. The obtained results show that SVM performed better than the two remaining classifiers.

In subsequent research, many more features have been tested by a large number of researchers (Kennedy and Inkpen, 2006, Kennedy and Inkpen, 2006, Yang et al., 2007, Ye et al., 2009, Ali et al., 2013). The state of the art of these features and methods can be found in the system presented in (Mohammad et al., 2013). Among submissions from 44 teams, the authors obtained the highest F1-scores for polarity classification in the 7th International Semantic Evaluation campaign SemEval 2013 (Nakov et al., 2013). The implemented system consisted in learning SVM using a variety of features such as: (i) the presence of word ngrams; (ii) the presence of character ngrams; (iii) the number of each Part Of Speech tag; (iv) the presence of positive and negative emoticons; (v) the number of elongated words (words with repeated characters); (vi) the number of words with all characters in upper case; (vii) the number of hashtags, etc. Moreover, the authors included the following features extracted from five English sentiment lexicons giving the valence of word

tokens: (i) the number of tokens expressing each sentiment class; (ii) the total score of the text document; (ii) the maximal score; and (iv) the score of the last token. Finally, they estimated the SVM complexity parameter by cross validation on the training set of the used benchmark.

Some studies used dependency trees in order to consider the syntactic relations between the words. (Matsumoto et al., 2005) extracted frequent word sub-sequences and dependency sub-trees and used them to construct features for an SVM classifier. Using, these features they improved the classification results of the basic ngrams. (Nakagawa et al., 2010) exploited the syntactic dependency structures of the sentences in text document classification. They proposed a dependency tree-based method for sentiment classification using Conditional Random Fields (CRFs) with hidden variables. In this method, the sentiment polarity of each dependency subtree, which is not observable in training data, is represented by a hidden variable. The polarity of the whole sentence is calculated in consideration of interactions between the hidden variables.

As presented before, most supervised sentiment classification works focus on feature engineering. The reason is that the performance of sentiment classifiers is strongly dependent on the choice of the feature representation. However, recent studies suggest to learn automatically word embeddings, which capture interesting and complex linguistic and semantic characteristics (last paragraph of section 2.3.2 page 20). (Socher et al., 2013) introduced Sentiment Treebank which uses fine grained sentiment labels to build recursive neural networks. Their system outperforms previous methods especially on negated phrases. (Tang et al., 2014) proposed a new sentiment specific word embeddings which encode sentiment information in the continuous representation of words. The authors used lists of positive and negative emoticons in order to collect large scale training corpora (10 million tweets). The learned sentiment specific word embeddings have been incorporated into an SVM classifier using the min, average and max convolutional layers to obtain the tweet representation (Collobert et al., 2011). They obtained comparable results with the top-performing systems using hand-crafted features. Among 44 teams, their system were ranked 2nd in SemEval 2014 (Rosenthal et al., 2014). Deep Learning based systems were also well ranked in the next evaluation campaigns SemEval 2015 (Rosenthal et al., 2015) and SemEval 2016 (Nakov et al., 2016). However, this kind of approaches require high computational resources in order to learn the word embeddings.

In this thesis, as it will be described in chapter 6, we tested various features and methods described above in the context of French Sentiment Classification. Several types of French texts have considered using the benchmarks described in the next section.

2.4.3 Benchmarks

A huge number of labeled data for sentiment classification have been released over the last few years. A common technique is to use labels that have been manually assigned on online web discussions. Many researchers (Turney, 2002, Cui et al., 2006) took advantage of websites where users provide ratings along with their reviews (IMDB, Amazon, Epinions, etc.). Others scrawled Twitter with a list of emoticons (Pak and Paroubek, 2010) or lists of positive and negative word hashtags (Mohammad and Kiritchenko, 2015). Finally, manual annotations can be performed to obtain datasets of higher quality (Wiebe et al., 2005). The same documents are usually annotated by many annotators in order to compute agreements measures. An exhaustive list of English sentiment classification benchmarks has been reported in (Pang and Lee, 2008). Here we present French benchmarks that have been studied in the case of sentiment-related challenges.

DEFT[4] is a French text mining challenge that evaluates methods and systems related to text mining. The third edition of this challenge (DEFT'07) concerned the classification of text documents according to their polarity (Grouin et al., 2009). The eleventh edition of the same challenge (DEFT'15) also concerned sentiment classification. The participants were asked to classify tweets according to their polarity, subjectivity and expressed emotions (Hamon et al., 2015). Table 2.3 describes the nature and the subject of the used benchmarks. These benchmarks are available publicly on the challenge website[5].

Table 2.3: The benchmarks used in sentiment-related French challenges.

Benchmark	**Description**	**Task(s)**	**Challenge**
See & Read	Movie, book and show reviews from the avoir-alire website[6]	Polarity	DEFT'07
Parliamentary Debate	Debate reports in the French National Assembly between 2002 and 2007[7]	Polarity	DEFT'07
Videos Games	Video game reviews from the jeux-videos website[8]	Polarity	DEFT'07
Climate	Tweets about climate change annotated under the ucomp project[9]	Polarity, subjectivity and emotion	DEFT'15

[4]DEFT: Défi de Fouille de Text
[5]deft.limsi.fr
[6]www.avoir-alire.com
[7]www.assemblee-nationale.fr/12/debats

Three benchmarks have been released for DEFT'07. On the one hand, See & Read and Videos Games associate product reviews with three polarities: good, medium and bad. On the other hand, Parliamentary Debate contains speech reports of parliamentarians who are for or against a given law. Therefore, this third benchmark associates its text documents with two polarities: for and against. Figure 2.3 shows the number of text documents for each polarity class in these three first benchmarks.

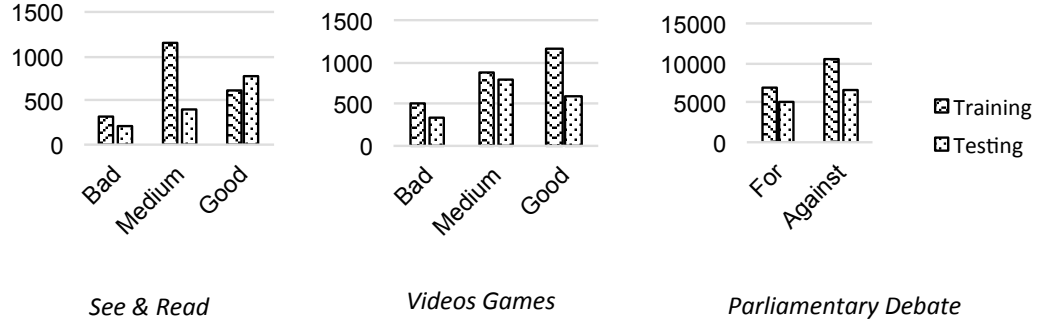


Figure 2.3: The number of text documents for each polarity class in the three benchmarks of DEFT'07.

The fourth benchmark (Climate) considers three sentiment classification tasks: (i) classifying tweets according to their polarity (positive, negative and neutral); (ii) classifying tweets according to their generic subjectivity class (information, sentiment, opinion and emotion); and (iii) classifying tweets according to their specific opinion, sentiment or emotion class (18 classes). Figure 2.4 shows the number of tweets in the classes defined in each classification task. For better visualization, the number of tweets is shown in logarithmic scale (base 10) for the second and the third tasks.

Regarding the documents length, Table 2.4 presents the average number of words per document for each benchmark. It appears that Videos Games has the longest text documents with more than one thousand words per document (on average). See & Read and Parliamentary Debate have long text documents (hundreds of words per document). Finally, the Climate benchmark has very few words per document, since tweets contain at most 140 characters.

The presented benchmarks will be used in this thesis in order to evaluate different combination of features, methods and lexicons for French sentiment classification. Most of them are extracted from Social Media (product reviews and tweets). However, it is still interesting to evaluate our methods on text documents from a different source (such as parliamentary debate reports).

[8] www.jeuxvideo.com
[9] www.ucomp.eu

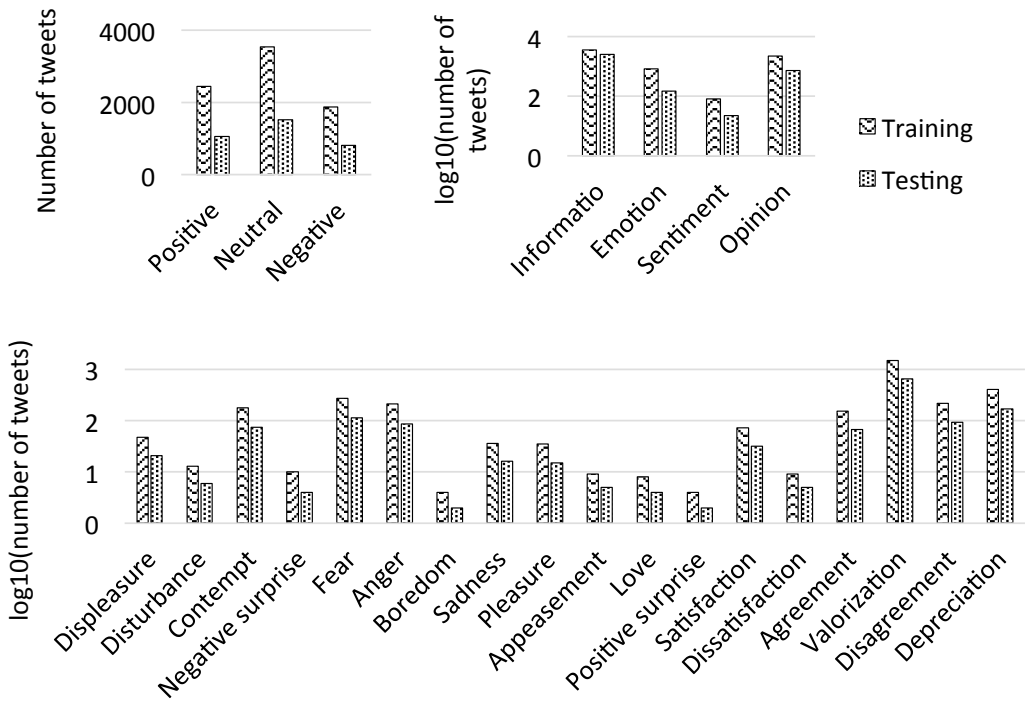


Figure 2.4: The distribution of tweets in each class of the DEFT'15 benchmark.

Table 2.4: The average number of words per document in each benchmark.

Benchmark	**Number of words**
See & Read	381
Videos Games	1,215
Parliamentary Debate	220
Climate	17

2.5 Conclusion

In this chapter, we presented some research paths in Social Media Mining that answer the questions raised in this thesis (namely who? and how?). The main distinction was made between network-based approaches and content-based ones. If most research in Social Media Mining use the network structure, we believe that the textual content may reveal complementary knowledge that could not be observed through the network analysis. Therefore, more attention has been given to the presentation of Text Mining techniques in Social Media. Finally, Sentiment Analysis which constitutes an example of Text Mining in Social Media has been detailed in the last section.

Network analysis in Social Media represents the data using graph theory (oriented graphs, multi-graphs, etc.). Our study was divided into two parts: (i) the extraction of these graph structures and (ii) the analysis of the extracted networks. First, we showed that some social data sources (such as online forums) can not be represented directly using graph theory. Indeed, in the literature, some heuristics have been proposed in order to extract social networks from these websites. In this thesis, we develop and evaluate our own rule-based network extraction heuristic, which is in part inspired from the existing ones. In fact, we define new rules such as: grouped posts, activator posts, questions, answers, etc. Then, we presented methods that use the extracted social networks in order to mine the importance and the expertise of the users. Many of these studies took advantage of the work done in ranking web pages according to their importance. In the following chapters, we propose a new measure inspired from these studies in order to predict the expertise of online forum users. Instead of taking into account only positive links, we propose to consider both positive and negative replies to a given user.

Then, we presented Text Mining in more details since the methods developed in this thesis are mainly based of the textual contents. First, we presented linguistic pre-processing steps that are usually applied to deal with Social Media texts. Then, we described text representations that prepare the unstructured text documents to the application of Machine Learning methods. Finally, we focused on Text Categorization which consists in classifying text documents to pre-established classes using annotated training data. We summarized the used classification models, evaluation metrics and annotation methods. These steps are essential for us in order to mine the textual Social Media contents. For instance, in the next chapter we will use Text Categorization in order to predict the author expertise of each forum post.

In the last section, we studied the task of Sentiment Analysis which allows the evaluation of text documents according to the expressed sentiments and opinions. This task has been detailed since it consists the second part of this thesis. First, we conducted a review of existing sentiment lexicons. We found that most of them have been proposed for the polarity of English terms. Therefore, in this thesis we describe the construction of a new French sentiment and emotion lexicon by translating and expanding the English NRC EmoLex lexicon. Then, methods from two main sentiment classification techniques have been discussed. The unsupervised methods which are based on sentiment lexicons and the supervised ones which use Text Categorization have been described. We mentioned that supervised approaches usually obtain better results. However, most of them dealt with English text documents. Therefore, by the end of this thesis we will present the application of the state of the art features, methods and resources on the French sentiment benchmarks that has been presented at the end of this chapter.

Part I

Expertise and Reputation

Chapter

3 Predicting User Expertise

Contents

3.1 Introduction 40
3.2 Materials and Methods 41
3.2.1 Corpora 41
3.2.2 Pre-processings 42
3.2.3 Annotations 43
3.2.4 Classification 44
3.3 Experiments 45
3.3.1 Cross Validation 45
3.3.2 Training and Testing on Different Data sets 46
3.4 Discussions 47
3.4.1 Results Interpretation 47
3.4.2 Error Analysis 48
3.5 Conclusion 49

3.1 Introduction

In this first work, we suggest to predict the expertise of the author who post in an online forum. Identifying expert posts is an important issue since it facilitates the access to the best answers that are more likely to be trustworthy and informative. The main objective of this first work is to use posts from websites, in which the expertise is indicated, in order to build efficient classification models that can predict the potential expertise in other forums. Therefore, the proposed method uses supervised Machine Learning algorithms in order to perform text categorization. In this work, health forums have been used in order to predict the medical expertise in forum posts.

Health forums are increasingly visited by both sick and healthy users when they want to get help and information related to their health. According to a study conducted by the Health On the Net foundation[1], 50% of e-patients use online health forums to acquire medical information. However, these forums are not limited to patients. More and more frequently, a significant number of medical experts are involved in online discussions. Indeed, some medical websites hire health experts (physicians, medical students, volunteers, etc.) and indicate explicitly their role. Others visit health forums unofficially and answer the patients questions without a special indication about their expertise. Being experts, they are able to clearly explain the problems, the symptoms, to correct false affirmations and to give precise and trustworthy answers. Furthermore, patients may acquire expertise through their own experience with a particular disease. After recovery, some of them go back to online forums in order to share their experience and help other patients.

The aim of this work is to distinguish between posts written by medical experts (either health practitioner or experienced patients) and by non-experts users. Due to the availability of annotated posts (forums that indicate the medical expertise explicitly on the website), supervised learning can be used in order to perform this task. We suggest to tackle this question by analyzing the posts content. A huge effort has been made in developing content-based classification systems for author profiling (Rangel et al., 2015). The idea is to discover unknown characteristics of the authors of a given text using supervised learning. Most studies concerned the discovering of age and gender from blogs, for which multiple features and classification models have been evaluated (Weren et al., 2014). The used classification models are the classical ones (SVM, NaiveBayes, Decision Trees, etc.), while the used features depend on the classification task (number of emotion terms, document length, etc.). Here, we focus on those that may be useful for medical expertise categorization.

First, (Tapi Nzali et al., 2015) showed that medical experts and patients use different vocabularies. Non-experts write more about symptoms and about themselves: (e.g. *"I have a headache"*), while experts should write more about treatments and about the patients: ('e.g. *"you should pass a mammography test"*). Therefore, a bag of words configuration along with a feature subset selection step are considered

[1] www.healthonnet.org

to capture these differences in the used vocabularies. Then, (Rangel and Rosso, 2016) studied the impact of emotions and sentiments in author profiling (age and gender). They proposed an emotion graph to model the way people use the language and the emotions when writing. They obtained respectively the first and the second best results for age and gender on the Spanish partition of PAN 2013 corpus. The use of emotions may also be interesting in our task since patients should use more emotions than the medical experts (for example to express the pain caused by the illness). Finally, (Grabar et al., 2016) compared documents written by medical doctors and researchers (clinical reports and scientific literature) with the patient discourse (discussions from health forums). They observed differences in the use of descriptors like uncertainty markers, non-lexical (smileys, repeated signs, etc.) and lexical emotional markers, and medical terms related to disorders, medications and procedures. In this work, all these features are considered in order to evaluate the most representative components of a forum post that allow to perform efficiently medical expertise categorization of forum posts.

The remainder of this chapter is organized as follows. First, the used health forums and the different steps of the proposed method are introduced in section 3.2). Then, section 3.3 presents the results of the conducted experiments and section 3.4 discusses them. Finally, section 3.5 concludes and gives some perspectives.

3.2 Materials and Methods

This section discusses the proposed method which consists in: (i) corpora acquisition; (ii) linguistic pre-processing; (iii) annotation of useful descriptors for medical expertise categorization and (iv) supervised classification (feature construction and model learning).

3.2.1 Corpora

Two French corpora have been collected from two health forums as described below.

AlloDocteurs.fr

A French health forum covering a large number of topics related to health such as alcoholism, pregnancy, sexuality, etc. 16,000 messages posted from June 2009 to November 2013 have been collected. The forum contains both expert and non-expert users. Medical experts include professional physicians and medical students. Even if their number is limited (16 medical experts over more than 6,000 registered users), their participation in the forum exchanges is important. Indeed, they posted more than 3,000 posts among the 16,000.

MaSanteNet.com

An online ask the doctor service that allows users to submit one or more questions to two doctors. The range of topics covered is also large. Users can ask questions on more than 20 different topics such as nutrition, dermatology, pregnancy, etc. More than 12,000 messages posted from January 2011 to March 2014 have been collected from this website. All the questions published on the website have answers. Therefore, the collected posts are equitably divided between patients' questions and doctors' answers.

Once collected, we applied a cleaning step to each website. First, very short posts (less than 10 characters) have been filtered out. Then, we removed quotes and author signatures that have been introduced inside the post content. Figure 3.1 presents the number of posts and words in the obtained data sets. On the one hand, it appears that the first corpus has fewer posts than the second one: approximately 4,400 posts for AlloDocteurs and approximately 12,000 posts for MaSanteNet. On the other hand, it appears that in both data sets, posts written by non-experts are usually longer than those written by medical experts.



Figure 3.1: The number of documents and the average number of words per document in each corpus.

3.2.2 Pre-processings

The following pre-processing steps are applied:

- **Slang Replacement:** Some abbreviations are frequently used in Social Media. Using a pre-established list, they have been replaced by the corresponding standard text (e.g. *lol* is replaced by *lot of laugh*);

- **User Tags:** User tags (e.g. *@Diana*) are identified in our corpora and replaced by the word *tag*;

- **Hyperlinks and Emails:** Hypertext links are replaced by the word *link* and email addresses by the word *mail*;

- **Pseudonyms:** Medical expert pseudonyms previously extracted from each website, are used to replace all their apparitions inside the posts by the word *fdoctor*. Similarly, pseudonyms of non-experts are extracted and used for their replacement by the word *fpatient*;

- **Lowercasing and Spelling Correction:** All words are lowercased and processed with the Aspell. The default Aspell French dictionary has been expanded with all the pseudonyms and all the medical words extracted from the used corpora. The medical terms are obtained after an annotation step as described in the next section 3.2.3.

3.2.3 Annotations

In order to categorize the discourse of medical experts and the discourse of non-experts, the descriptors proposed in (Grabar et al., 2016) have been annotated using the Ogmios platform (Hamon and Nazarenko, 2008). This annotation step allows us to include them easily as features in the classification step.

- **Medical Concepts:** Terms belonging to three semantic types (diseases, treatments and procedures) are detected using the following medical resources: the Systematized Nomenclature of Human and Veterinary Medicine[2], the Thériaque database[3], the Unified Medical Language System[4];

- **Emotions:** A French emotion lexicon has been used to annotate terms conveying six types of emotions (joy, surprise, fear, sadness, anger and disgust). The lexicon contains about 14,000 emotional terms. The compilation process of this lexicon will be presented in chapter 5. In addition to this lexicon, some non-lexical expressions of emotions such as: repeated letters, repeated punctuation signs, smileys, slang and capital letters are detected and annotated with specifically designed regular expressions;

- **Uncertainty:** A set of uncertainty words (Grabar et al., 2016) is used to annotate terms conveying uncertainty meaning (for example: to seem, possible, probably, etc.). Three levels of uncertainty are considered: weak, medium and strong.

[2]www.ihtsdo.org/snomed-ct
[3]www.theriaque.org
[4]www.nlm.nih.gov/research/umls

3.2.4 Classification

As mentioned before, the proposed approach is based on supervised Machine Learning. Here we describe the implemented features and the used classification models.

Feature Construction

In addition to the features based on the annotation step, the number of misspellings and question marks are included in the categorization task. Figure 3.2 shows the number of medical concepts, emotions terms, uncertainty markers, misspellings and question marks in each benchmark. It appears that non-experts use more medical concepts and emotion terms, ask much more questions and do more spelling mistakes, while medical experts use slightly more uncertainty markers (usually to make an uncertain diagnosis). Therefore, 14 attributes representing these descriptors are included in our classification task (medical concepts: three attributes, emotion terms: six attributes, uncertainty markers: three attributes, questions: one attribute, misspellings: one attribute). For each attribute, we compute the number of occurrences normalized by the corresponding post length. The length of each post corresponds to the number of words it contains. We call these features "Dictionary-Based Features".

Moreover, a bag of words representation is considered. Words that appear at least two times in the training sets are included. Each word is represented by its normalized number of occurrences (number of occurrences divided by the corresponding post length).

Feature Selection

Feature subset selection is applied to select the most discriminant features: those that frequently appear in only one category of posts. Therefore, the selected features should characterize one category of users. Feature selection is known to reduce the feature dimensionality and improve the classification results. In our case, the Information Gain method is used as a filter to select attributes in each experiment.

Classification Models

The Weka Data Mining platform (Hall et al., 2009) is used to learn the classification models. Since feature selection does not remove redundant attributes, models that assume the independence of the features (such as Naive Bayes) are not adapted. In this work, we tested the following models: Support Vector Machines (SVM SMO), decision trees (J48 and Random Forest) and rule based models (JRip). The Weka default configuration is used train each classification model.

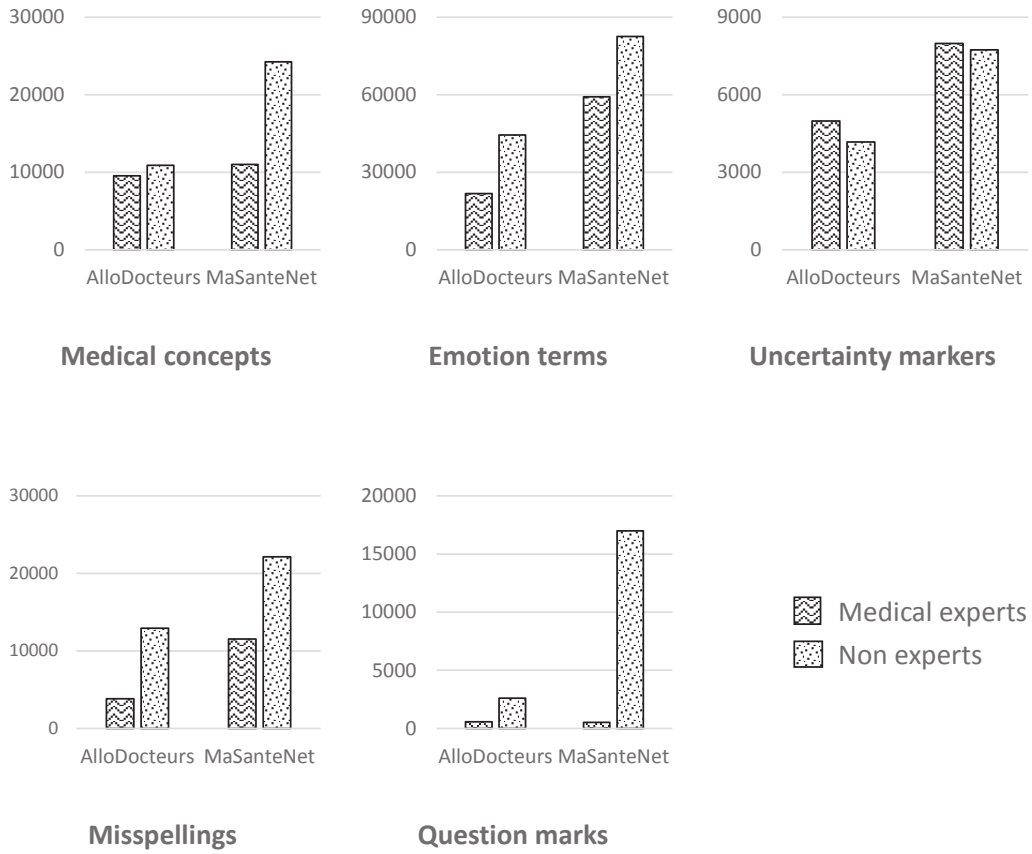


Figure 3.2: The number of medical concepts, emotion terms, uncertainty markers, misspellings and question marks in each corpus.

3.3 Experiments

In this section, we evaluate different combinations of features with the above mentioned classification models using weighted F1-measures. For a balanced data set, chance will produce a weighted F1-measure of 0.5 that can be considered as a baseline for evaluating our results.

3.3.1 Cross Validation

First, 10-fold cross validation has been performed on each data set separately. The features construction, selection and classification models are learned on the training subset of each fold. Moreover, the same training and testing sets are used to learn and test our four classification models in each fold.

Tables 3.1 and 3.2 show that on both data sets, bag of words induce high weighted F1-measures. They obtain more than 90% on AlloDocteurs and perfect classification F1-measures (100%) on MaSanteNet. On the other hand, the dictionary-based markers induce lower weighted F1-measures: between 71% and 75% on AlloDocteurs

Table 3.1: The weighted F1-measures obtained with 10-fold cross validation on AlloDocteurs.

Feature group	**SVM SMO**	**J48**	**Random Forest**	**JRip**
Bag Of Words	92	90.6	92.1	89.7
Dictionary-Based Markers	71.6	73	74	75
Bag Of Words + Dictionary-Based Markers	**92.7**	90.7	**92.7**	90.3

Table 3.2: The weighted F1-measures obtained with 10-fold cross validation on MaSanteNet.

Feature group	**SVM SMO**	**J48**	**Random Forest**	**JRip**
Bag Of Words	**100**	**100**	**100**	**100**
Dictionary-Based Markers	88.9	91.6	93.6	92
Bag Of Words + Dictionary-Based Markers	**100**	**100**	**100**	**100**

and between 88% and 94% on MaSanteNet. Regarding the classification models, SVM SMO and Random Forest obtained the highest F1-measures on MaSanteNet. Finally, the use of the dictionary-based features along with the bag of words configuration does not change the results (the obtained F1-measures are almost the same as those obtained only with bag of words). The presented results may indicate that our models are dependent on the forum used for learning. Therefore, we evaluate the genericity of the models learned on each forum and test them on the other forum.

3.3.2 Training and Testing on Different Data sets

In this work, we assume that models learned on specific forums can be used efficiently on other forums. In order to evaluate this claim, two more experiments are conducted. In each experiment, features and classification models are constructed and learned on one data set and tested on the other data set.

Tables 3.3 shows that models learned on AlloDocteurs obtain significantly high F1-measures. The bag of words used alone or with the dictionary-based features induce more than 96% in terms of weighted F1-measures when tested on MaSanteNet. Once again, Random Forest obtains the highest F1-measure. The dictionary-based features induce F1-measures between 57% and 70%. These results show that the models learned on AlloDocteurs remain highly efficient when applied on MaSanteNet. However, Table 3.4 shows that the classification models learned on MaSanteNet obtain low F1-measures. The weighted F1-measures of the bag of words

Table 3.3: The weighted F1-measures obtained with AlloDocteurs as training set and MaSanteNet as testing set.

Feature group	**SVM SMO**	**J48**	**Random Forest**	**JRip**
Bag Of Words	96.6	97.7	98	96.9
Dictionary-Based Markers	57	62.1	69.6	69.6
Bag Of Words + Dictionary-Based Markers	96	97.3	**98.2**	96.6

Table 3.4: The weighted F1-measures obtained with MaSanteNet as training set and AlloDocteurs as testing set.

Feature group	**SVM SMO**	**J48**	**Random Forest**	**JRip**
Bag Of Words	37.3	33.3	46.3	33.3
Dictionary-Based Markers	**57.1**	52.9	53.2	55.3
Bag Of Words + Dictionary-Based Markers	37.5	33.3	43.7	33.3

features used alone or with the dictionary-based features drop significantly when tested on AlloDocteurs (between 33% and 44%). Since our benchmarks have balanced classes, chance produces better F1-measures than the learned bag of words classifiers (50%). The weighted F1-measures obtained by the dictionary based features drop slightly when tested on AlloDocteurs (between 52% and 58%). SVM SMO induces the highest F1-measure using these features. Finally, we can conclude that the bag of words models learned on MaSanteNet are extremely context dependent, which make this forum inappropriate for training generic models.

3.4 Discussions

Here we discuss the obtained results and present an error analysis study.

3.4.1 Results Interpretation

Despite the high F1-measures obtained with cross validations on both data sets, the models learned on AlloDocteurs remain efficient when applied on MaSanteNet. However, those learned on MaSanteNet gave lower F1-measures when applied on AlloDocteurs. These results can be explained by the fact that the first website is a health forum, in which 16 medical experts participate in the forum discussions. They post messages in any thread where their expertise is needed, which make the

discourse of the medical experts more extensive and diversified. Therefore, models learned on this forum may cover the topics and the discourse found on MaSanteNet. On the other hand, MaSanteNet is a limited health forum (an Ask the doctor service) in which only two medical experts answer all questions. There is no long discussions since each thread contains only one question and one answer. The answers are formed following the same pattern, which makes the discourse of the medical experts very specific to this website. For this reason, MaSanteNet appears to be less suitable for learning classification models that can be used on other forums.

Using emotions, uncertainty markers and medical concepts, (Grabar et al., 2016) obtained F1-measures between 91% and 95% when classifying forums posts produced by patients and clinical reports produced by medical experts. This work shows the limits of using these markers in categorizing the patients discourse and the medical experts discourse when the text documents are of the same nature (forum posts). Our results suggest to use bag of words features, which are the most adapted to perform such categorization. This result confirms those obtained in the author profiling challenge PAN (Rangel et al., 2015), where the best systems used content-based features (bag of words, TF-IDF n-grams, etc.).

3.4.2 Error Analysis

An error analysis of the 10-fold cross validation applied on AlloDocteurs has been performed. In each fold, our four classification algorithms have been trained on 90% of the data using all the features (bag of words and dictionary-based markers) and tested on the remaining 10%. If at least three algorithms agree to classify a post to the wrong category (with respect to the role given on the website), the post is to be studied manually. Figure 3.3 shows the percentage of posts correctly and incorrectly classified by a majority vote between the four trained models. We study 164 posts (4%) among which 107 were written by patients but classified as medical experts (3%) and 57 which were written by medical experts but classified as patients (1%).

On the one hand, the manual analysis of the 107 posts classified as medical experts allowed us to find new users having medical expertise that is not indicated on the website. They may be either medical physicians (e.g. *"many similar cases come to see us in the hospital"*) or only users that had the same experience before (e.g. *"the pain will disappear in few days, my mother had the same surgery"*). These users posted 79 messages among the 107, which confirms that medical experts may participate in the discussions even if their role is not explicitly indicated. In this case, only 47 posts have been considered as truly misclassified.

On the other hand, the manual analysis of the 57 posts that has been written by medical experts and classified as patients showed that medical experts may have the same discourse as patient (e.g. they may ask questions). This observation highlights that even medical experts may lack of expertise in a particular topic or need precision on the patient condition.

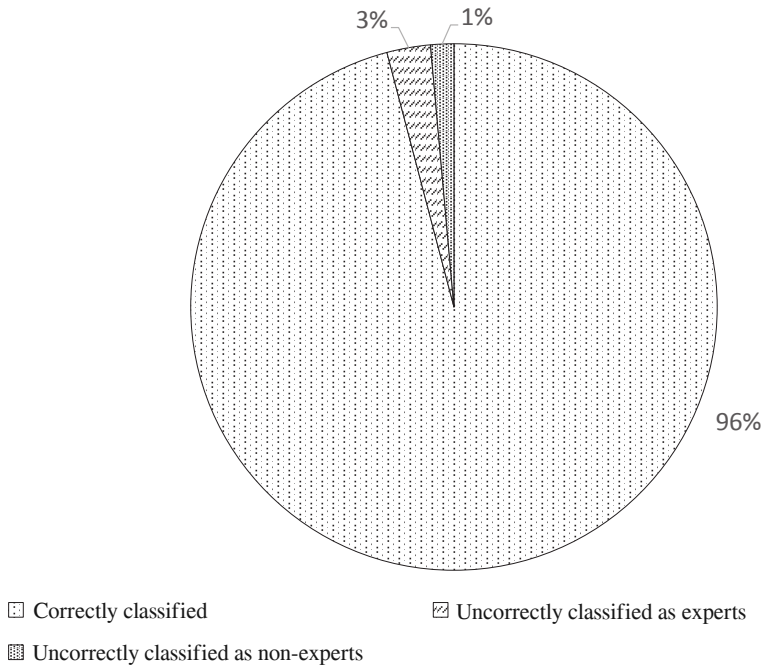


Figure 3.3: The percentage of posts correctly and incorrectly classified by a majority vote using 10-fold cross validation on AlloDocteurs.

3.5 Conclusion

In this chapter, we presented a supervised learning approach designed to distinguish posts written by medical experts from those written by non-experts in French health forums. The proposed approach uses forums that hire medical experts and indicate explicitly their role in order to learn classification models that can predict the medical expertise on other health forums. The conducted experiments confirm that models learned on appropriate forums where many medical experts participate in various discussions can be applied on other websites with satisfactory results. Regarding the most adequate features to this task, the obtained results show very high F1-measures with the bag of words features (up to 98% using Random Forrest). The dictionary-based features used in (Grabar et al., 2016) show low F1-measures. Even if these features have been efficient in categorizing the patients discourse present in forums and the medical experts discourse present in clinical and reports and scientific literature, this work shows their limits when the text documents are of the same nature (forum posts). Analyzing the misclassified posts allowed us to find out that medical experts may write posts in online health forums even if their medical role is not indicated on the website. The study of the misclassified posts also showed that the expertise of a user may change according to the discussed topic.

Direct extensions of this first work may be studied for the short run. On the one hand, it would be interesting to track the evolution of the user expertise over time. Indeed, online forum users tend to acquire expertise over time. They usually start as information consumers (reading discussions, asking questions) to finish as information producers (posting informative answers). This observation has been reported in other Social Media services such as online reviews (McAuley and Leskovec, 2013). Since the method proposed in this chapter allows us to predict the expertise at the message level, we can easily trace the messages posted by each user and show the evolution of his expertise. On the other hand, it is wildly admitted that the user expertise may change according to the discussed topic (Guy et al., 2013, Omidvar et al., 2014). A user who has expertise in a given subject may be unskilled in another one. This observation has been noticed in this work, where some expert posts have been classified as non-experts. The study of these misclassified posts has shown that health professionals may lack of expertise in certain domains. Since, online forums are usually organized into specific categories, we may study the user expertise throw the messages he posts in each category.

The method presented in this first chapter deals with the expertise at the post level. However, it would be more interesting to infer the expertise at the user level. Online forums usually compute a reputation value for their users based on the votes they receive when posting useful answers. Indeed, many social websites allow their users to score (like, vote up, etc.) posts according to their correctness and usefulness. These scores are then aggregated to compute a reputation value for each user. However, explicit rating functionalities are rarely used in many online communities and especially in health forums. Indeed, users of these communities prefer posting a new message where they express their agreement or thanking rather than pressing a like button. Therefore, we believe that the user reputation may be inferred by analyzing the textual content of the posts addressed to him. The goal is to detect positive (agreement and valorization) and negative (disagreement and depreciation) replies and aggregate them to compute a user reputation value. In the next chapter, we will explore this idea by developing and evaluating a complete content based-method.

Chapter

4 Computing User Reputation

Contents

4.1 Introduction 52

4.2 Materials and Methods 53

4.2.1 Theoretical Framework 53

4.2.2 Corpora 54

4.2.3 Extracting the Interaction Network 56

4.2.4 Predicting Positive, Negative and Neutral Replies 57

4.2.5 Proposed Metrics 58

4.3 Evaluations and Discussions 59

4.3.1 Evaluating the Network Extraction Step 59

4.3.2 Evaluating the Trust Prediction Step 60

4.3.3 Evaluating the Proposed Metric 61

4.4 Conclusion 63

4.1 Introduction

Online forums are increasingly visited to share, to discuss and to get information and help for all aspects of our lives. They are areas of exchange generated by their own users. Therefore, the veracity and the quality of the posted information vary according to the expertise of their author. With the massive and rapid growth of these conversational social spaces, it becomes very difficult for human moderators to separate good posts from bad ones. Consequently, more and more forums are implementing automated trust and reputation metrics to infer the trustworthiness of posts and the reputation of their authors. These metrics vary from ranks based on a simple post count to more elaborated reputation systems based on collaborative ratings. If the first category of metrics tries simply to reward users according to the number of posts, the second category uses collaborative intelligence to rate the user's posts and then aggregate these ratings to give him a reputation value (Lampe and Resnick, 2004). This idea has been successfully applied in many online forums such as news groups (Slashdot[1]), question-answering websites (Stack Exchange[2]), etc. The communities behind these forums are usually computer scientists, programmers, gamers or simply users interested in technical issues. However, collaborative rating is not popular in other communities such as health forums, where users prefer to post a new message in order to thank each other rather than clicking the "like" or "vote up" button. The objective of this work is to use this implicit collaborative intelligence hidden in the textual content of the replies in order to infer user reputations.

(Wanas et al., 2008) proposed a method inspired from forums that use collaborative intelligence to rate posts. This method automatically scores posts based on their textual content. The authors tried to model how users would perceive a post as good or as bad. Indeed, five categories of features have been considered: (i) Relevance features: reflect the appropriateness of a post to the sub-forum topic or to the thread topic; (ii) Originality features: measure the amount of new knowledge brought by a post; (iii) Forum specific features: include the number of text quotations and post replies; (iv) Surface features: compute the time difference between posts, the length of posts and the formatting quality such as the number of smiley and capital letters used; and finally (v) Posting component features: consider the use of web links and questions. Natural language processing techniques have been used to compute these features. For example, the relevance of a post to the thread topic has been computed as the percentage of the post's words that appear in the topic title and the leading posts. The originality of a post has been computed as the lack of lexical similarity between this post and the previous posts, etc. The authors experimented their approach on 200 threads from the Slashdot online forum. Using different combinations of the above mentioned features, they trained classification models and compared their results with rated scores obtained from Slashdot. Unlike (Wanas et al., 2008), we believe that the reputation of a given user is defined by the

[1] www.slashdot.org
[2] www.stackexchange.com

trust expressed by other users. Therefore, instead of inferring the user's reputation from his own posts, we suggest to consider the messages addressed to him.

Many definitions of trust and computational trust exist in the literature (Deutsch, 1962, Marsh, 1994, Sztompka, 1999, Golbeck, 2009). Here we define the trust that a user A has in another user B as: *"the belief of A in the veracity of the information posted by B"*, and the reputation of a user A as *"the aggregation of trust values given to user A"*. To infer such trust from textual replies and aggregate user reputations, we need to know both the recipient of each forum message and the trust expressed in it. However, the forum structure does not always provide explicit quoting or direct answering functionalities. Besides, when these functionalities are provided, many users prefer posting a message answering the whole thread rather than a one answering or quoting another specific message. In order to deal with this issue, we propose a rule based heuristic to extract an interaction network where the nodes are the users and the edges are the replying posts. Regarding the semantic evaluation of each post's content, the features that we are looking for are agreement and valorization for trust, and disagreement and depreciation for distrust. The rest of posts are considered as neutral. Finally, we propose a metric to aggregate trust and distrust replies that a user receives and infer his reputation in the forum. The proposed reputation metric considers propagation aspects by giving more weight to the replies posted by trusted users and less to the replies posted by untrusted ones. Manual annotations have been performed in order to evaluate the results of the proposed approach.

The rest of this chapter is organized as follows. Section 4.2 presents the theoretical framework, the used forums and the proposed methods. Section 4.3 presents and discusses the obtained results using manual annotation. Finally, section 4.4 concludes and gives our main perspectives.

4.2 Materials and Methods

In this section, we briefly describe the theoretical framework, the used forums and the proposed methods which are divided into three main steps: (i) the extraction of the interaction network; (ii) the detection of positive and and negative replies, and (iii) the aggregation of these replies in order to compute the user's reputation.

4.2.1 Theoretical Framework

In our case, a specific user may reply many times to another user (see replies from v_1 to v_2 in Figure 4.1). Therefore, we define a directed multigraph to model the interactions; $G = (V, E, t, r)$ where:

V is the set of users.

E is the multiset of 'reply-to' edges between these users.

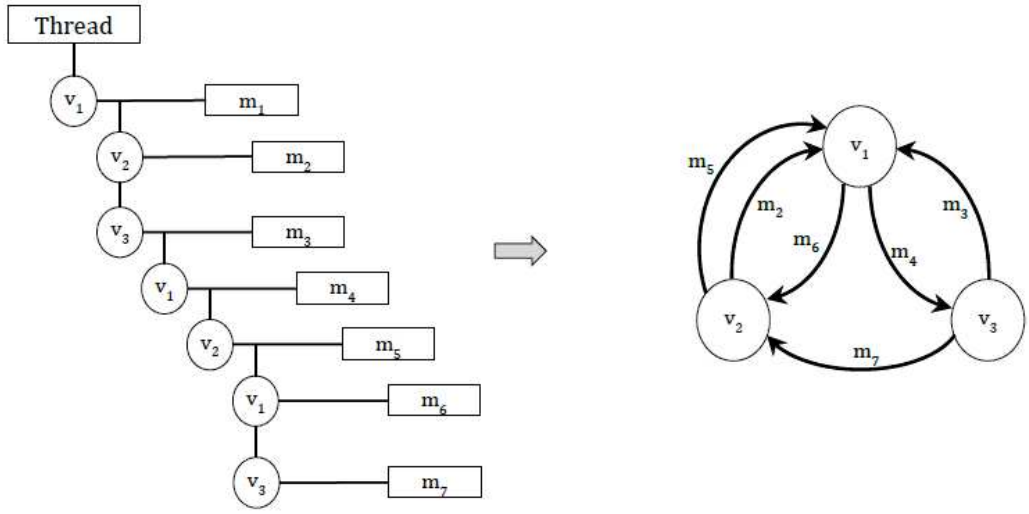


Figure 4.1: The representation of a thread discussion using multigraph where nodes represent users (v_i) and edges represent replies (m_i).

t is a function that returns the transmitter of a reply defined as follows:

$$\begin{array}{rccc} t & : & E & \rightarrow & V \\ & & e & \mapsto & t(e) \end{array}$$

r is a function that returns the recipient of a reply defined as follows:

$$\begin{array}{rccc} r & : & E & \rightarrow & V \\ & & e & \mapsto & r(e) \end{array}$$

Let $v \in V$ be a user. Then $E_v \subset E$ is the set of edges that reply to the user v:

$$E_v = \{e \in E; r(e) = v\}$$

Let E_v^+, E_v^- and $E_v^n \subset E_v$ be the subsets of positive, negative and neutral edges that reply to the user v. Note that:

$$E_v = E_v^+ \cup E_v^- \cup E_v^n$$

$$E_v^+ \cap E_v^- = E_v^+ \cap E_v^n = E_v^- \cap E_v^n = E_v^+ \cap E_v^- \cap E_v^n = \emptyset$$

4.2.2 Corpora

Again, two French health forums have been collected.

CancerDuSein.org

A French health forum specialized in breast cancer. It gathers women with breast cancer or their affected families. 1,050 threads have been collected which amounts 16,961 messages posted by 675 users. It represents all the data that have been posted between October 2011 and November 2013. Some threads have more than 500 posts, which make the use of semi-automatic systems a challenging task. This forum allows users to thank each other using a "like" button at the bottom of each post, but this functionally is rarely used. Indeed, less than 1.4% of messages received at least one "like". On the other hand, CancerDuSein.org gives a rank to each user based on the number of posts since his registration as presented in Table 4.1. However, we believe that these ranks are not sufficient to infer reputations. Indeed, users may post plenty of empty or useless messages in order to obtain better ranks.

Table 4.1: The number of users having each rank and the range of postings to acquire it in CancerDuSein.org.

Rank	**Rang of postings**	**Number of users**
New member	$[0, 20[$	564
Regular member	$[20, 40[$	42
Accustomed member	$[40, 80[$	26
Active member	> 80	41
Moderators	-	2
Total	-	**675**

Forum-thyroide.net

A French health forum dealing with thyroid dysfunctions. 269,073 messages posted between April 2004 and March 2015 have been collected. These posts are organized into 37,857 threads and posted by 13,803 users. Table 4.2 presents the ranks and the number of users having each rank. The forum is managed by a moderator and 7 forum users who are members of the executive board fo the association. 14 users received the status of active members but this rank seems to be independent from the number of postings. Finally, 374 users are registered as donating members. In order to become a donating member, users have to email the executive board, fill-in specific forms and a contribution to the association fees. Therefore, this forum gives manual ranks to its users which are not based on the number of postings. Finally, Forum-thyroide.net also allows users to thank each other by voting up a given message but as in the first forum, this functionally has been used only few times. Indeed, users often post a new message where they express their thanking rather than clicking on this button.

Table 4.2: The number of users having each rank in Forum-thyroide.net.

Rank	Number of users
Regular user	12,407
Donating member	374
Active member	14
Member of the executive board	7
Moderator	1
Total	**13,803**

4.2.3 Extracting the Interaction Network

In order to extract the interaction network, we propose the following rule based heuristic. Some rules have been inspired from the literature (Gruzd and Haythornthwaite, 2008, Forestier et al., 2011) and the remaining ones have been proposed after manual annotation of discussions extracted from the used forums. The proposed rules are checked sequentially; i.e. if a message does not match the first rule the heuristic will check the second rule, if it does not match the second rule the heuristic will check the third rule, and so on. The first post in each thread does not reply to anybody.

1. **Explicit quoting:** CancerDuSein.org and Forum-thyroide.net allow their users to explicitly quote posts on the forums. Most of them have been detected automatically using the HTML tags and by comparing the content of these tags and the pseudonym of the quoted user with the messages posted before in the same thread. Few of them have been associated to the corresponding recipient manually because the quoted text has been modified or truncated by the user;

2. **Second posts:** Messages posted at the second place in each thread have been considered as replying to the first one;

3. **Names and pseudonyms:** If a message contains the pseudonym or the name of a user who previously posted in the same thread, then this user is considered as the recipient of the message. The pseudonyms have been extracted automatically while the names have been extracted from the signatures and validated manually. The following pre-processing steps have been applied to detect names and pseudonyms: (i) remove all non-alphabetic characters except spaces (***John Woe 34*** $\mapsto$ *John Woe*); (ii) Replace all accented characters with the corresponding non-accented ones (*Jérôme* $\mapsto$ *Jerome*); (iii) Lowercasing (*Julien* $\mapsto$ *julien*);

4. **Grouped posts:** If a message contains a group marker (“hello everyone”, “Hi

girls", "Thank you all", etc.) then users (up to three users) who previously posted in the same thread are considered as recipients for this post;

5. **Second person pronouns:** In French, singular second person pronouns and plural second person pronouns are different ("tu" and "vous"). If a singular second person pronoun is used then the recipient is considered to be the author of the previous post;

6. **Activator posts:** If the activator posts a new message in the same thread, we consider that his new message is addressed to the users who posted after his last message in the thread (up to three users);

7. **Questions:** If the message contains a question, then the message is addressed to the users who previously posted in the same thread (up to three users);

8. **Answers:** If a question has been posted before in the thread, the recipient is the author of this question;

9. **Default:** If none of the rules presented above is satisfied, we consider that the recipient is the author of the message.

4.2.4 Predicting Positive, Negative and Neutral Replies

Once the interaction network is constructed, we need to classify each post with one of the following three classes: (i) positive: the post expresses trust to its recipient; (ii) negative: the post expresses distrust to its recipient; (iii) neutral: otherwise.

Building Lists of Trust and Distrust Expressions

We manually created two lists of expressions that should indicate if a message expresses trust (or distrust) to the recipient. These lists have been obtained by manual annotations of a set of threads (20 threads from each website) using the brat tool[3]. The annotators were asked to choose trust, distrust or neutral for each thread post and to indicate the expressions that justify their choice. These expressions have been manually validated and automatically corrected, lowercased, lemmatized and enriched.

Handling Negation

If a trust expression is under the scope of a negation term, it is considered as a distrust expression and vice versa. The scope of a negation term may be two words after, two words before, or two words after and two words before according to the nature of the negation term.

[3]brat.nlplab.org

Computing the Frequencies and Classifying the Posts

All posts have been automatically lowercased, lemmatized and corrected using the Aspell spell checker. Then, posts containing more trust expressions than distrust ones have been classified as positives. Those containing more distrust expressions than trust ones have been classified as negative. The rest of posts have been classified as neutral.

4.2.5 Proposed Metrics

Authority based algorithms (such as PageRank) consider only the structure of the network to compute the user importance. They use the links without evaluating their semantics. In our case, we want to consider the above described positive and negative replies to compute the user reputation. Therefore, we propose the following measure inspired from PageRank.

For each user v, we define a reputation value $R(v)$ as follows:

$$R_{n+1}(v) = \begin{cases} \dfrac{\sum_{e \in E_v^+} R_n(t(e))}{\sum_{e \in E_v^+} R_n(t(e)) + \sum_{e \in E_v^-} R_n(t(e))} & , if E_v^n \neq E_v \\ 0.5 & , Otherwise \end{cases}$$

This equation is recursive and can be computed by starting with reputations equal to one and iterating until it converges. The proposed reputation equation depends on both the number of trust and distrust replies a user receives and the reputations of the users who posted these replies. Particularly, the more a user receives replies expressing trust and the less he receives replies expressing distrust, the better is his reputation. Furthermore, a reply expressing trust or distrust is weighted by the reputation of the user who posted it, such that replies posted by trusted users are given more weight.

On the other hand, the proposed reputation calculation does not take into account neutral replies, as they do not convey any trust nor distrust. However, the proportion of neutral replies a user receives may be an additional information to the computed reputation. Moreover, the number of replies a user receives gives an indication on the reliability of the computed reputation. The more replies he receives the more confidence we have in the computed reputation. In order to include these two additional observations, we define the following metrics:

$$NeutralRate(v) = \begin{cases} \dfrac{|E_v^n|}{|E_v|} & , if E_v \neq \emptyset \\ 0 & , Otherwise \end{cases}$$

$$Reliability(R(v)) = \begin{cases} \dfrac{|E_v|}{maxR} & , if |E_v| < maxR \\ 1 & , Otherwise \end{cases}$$

Where maxR is a constant that represents the maximum replies that a user should receive in order to have a reliability equal to 1 in his reputation.

4.3 Evaluations and Discussions

In order to evaluate the presented methods, we manually annotated 20 threads from each website as presented in table 4.3. Each thread has been annotated by three different annotators. The annotations performed on the first forum have been considered in the elaboration the above presented methods. Therefore, this forum can be considered as a development set. The annotations performed on the second forum have been used only for evaluating the systems. Therefore, this forum can be considered as a testing set.

Table 4.3: The number of annotated threads and messages from each website.

Forum	Number of threads	Number of messages
CancerDuSein.org	20	214
Forum-thyroide.net	20	141

4.3.1 Evaluating the Network Extraction Step

First, the annotation goal was to find manually the recipient(s) of each message (prior assesment). The annotators were asked to copy and past the pseudonym(s) of the author(s) to whom each message is addressed. 20 threads from each forum were annotated to test our rule based heuristic. Each thread has been annotated by three different annotators. Classical measures of agreement are not adapted to this situation. Here we simply present the percentage of links (message to recipient) found by all of the three annotators, those found by two out of the three annotators and those found by only one annotator. Table 4.4 shows that the majority of links were found by all the annotators (more than 53% on both forums). Between 25% and 29% have been found by two annotators. Finally, less than 19% have been found by only one annotator.

Using these annotations, the quality of the developed heuristic was evaluated. The links obtained automatically were compared with those obtained from the annotations by considering only those that have been validated by two or more annotators (a majority vote). We compare the results of our heuristic with two baselines. The first baseline considers the activator of the thread as the recipient of all the messages posted in this thread (Activator). The second baseline considers the author of the previous message as the recipient (Previous). Table 4.5 presents the obtained precision, recall and F1-measures.

It appears that the second baseline (Previous) outperforms the first one (Activator) on both forums. However, our heuristic obtained higher F1-measures than both baselines. Moreover, we notice that our heuristic remains efficient on the second forums (which contains threads unseen at the time of developing it). The

Table 4.4: The percentage of links found by one, two or three annotators in each forum.

Forum	Found by	Percentage of links
CancerDuSein.org	3 annotators	53.4%
	2 annotators	28.3%
	1 annotator	18.3%
Forum-thyroide.net	3 annotators	60.3%
	2 annotators	25.7%
	1 annotator	14%

Table 4.5: The evaluation of the network extraction heuristic on both forums.

Forum	Method	Precision	Recall	F1-measure
CancerDuSein.org	Baseline 1: Activator	0.52	0.33	0.40
	Baseline 2: Previous	0.65	0.42	0.51
	Heuristic	**0.81**	**0.66**	**0.73**
Forum-thyroide.net	Baseline 1: Activator	0.46	0.41	0.43
	Baseline 2: Previous	**0.79**	0.72	0.75
	Heuristic	0.72	**0.82**	**0.77**

second baseline (Previous) obtains higher precision on Forum-thyroide.net than the proposed heuristic. This observation may be explained by the fact that many rules consider more than one recipient (which is not the case for the baselines). Therefore, our heuristic has more chances of finding untagged or incorrect links.

4.3.2 Evaluating the Trust Prediction Step

The same threads have been also annotated with the purpose of detecting positive and negative replies. Again, each message has been annotated by three annotators. The obtained Fleiss' Kappa was equal to 0.61 for CancerDuSein.org and 0.69 for Forum-thyroide.net showing a substantial agreement between the annotators. The annotations have been combined using a majority vote. If each trust value (positive, negative and neutral) obtains exactly one annotation, we check again the message and chose the correct class by consensus.

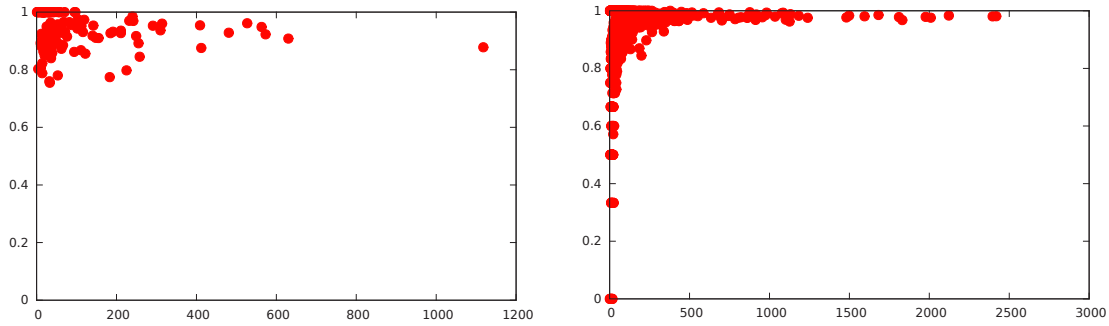
The obtained weighted average precisions, recalls and F1-measures obtained are presented Table 4.6. It shows that the automatic system based on trust and distrust expressions obtains more than 0.79 for all the evaluation metrics. Furthermore, it remains efficient on the second forums (which contains unseen messages by the compilation time of the used expressions).

Table 4.6: The evaluation of the trust prediction on both forums.

Forum	Precision	Recall	F1-measure
CancerDuSein.org	0.86	0.84	0.84
Forum-thyroide.net	0.82	0.79	0.81

4.3.3 Evaluating the Proposed Metric

In our experiments, the constant maxR has been fixed to the average number of replies received by the users. The reputations of the users having reliabilities greater than 0.5 are presented in figure 4.2. It shows the computed reputations of the considered users according to the number of posted messages in each forum.



CancerDuSein.org Forum-thyroide.net

Figure 4.2: The computed reputations according to the number of postings in each forum.

In both forums, the computed reputations are relatively high, which may be explained by the nature of the topics discussed in these forums. Indeed, little distrust is expressed in health forums since users aim at first to exchange emotional support. Furthermore, simply counting positive and negative expressions inside a text are usually biased, since users tend to use positive expressions even when they express a neutral or a positive meaning (Taboada et al., 2011). It would be interesting to compare this method with a learning one in order to better evaluate the trust prediction step.

Tables 4.7 and 4.8 present the average reputation of each user rank on each forum. As expected, the averaged reputations of all the ranks on both forums are high ($>$0.928). Then, the moderator of Forum-thyroide.net and the two moderators of CancerDuSein.org had perfect reputations (1). Indeed, the moderator of Forum-thyroide.net is a former patient who had thyroid cancer. After recovery, she decided to create this forum in order to help other patients. Therefore, she is often thanked by the forum users since her posts are of interest to them. Furthermore, it appears that the user reputation is independent from the number of posts. In CancerDuSein.org, the average reputation of the users having posted more than 80 messages

(the active members) is smaller than the average reputations of the users who posted less than 80 messages (the new members, the regular members and the accustomed members). On the other forum, the ranks are given by the forum moderator. They seem to be more adequate with our computed reputations.

Table 4.7: The average reputation of each user rank in CancerDuSein.org.

Rank	**Average reputation**
New member	0.949
Regular member	0.954
Accustomed member	0.946
Active member	0.928
Moderators	1

Table 4.8: The average reputation of each user rank in Forum-thyroide.net.

Rank	**Average reputation**
Regular user	0.941
Donating member	0.958
Active member	0.966
Member of the executive board	0.970
Moderator	1

Evaluating our computed reputations is a real issue. (Wanas et al., 2008) compared the computed scores with those appearing on Sloashdot. In our case, we can not find such gold data set since users who post a new message expressing thanking or agreement do not click on the like button. In order to better evaluate the computed reputations, we asked the moderator of Forum-thyroide.net a list of users who should have the best reputations on her forum. She opened a private thread with the active members and the executive board of the associated. After discussion, they chose eight users among them who should have the best reputations. When compared with our computed reputation values, these users obtained values greater than 0.98 and all of them had a reliability equal to 1. The pointed eight users are ranked among the 60 highest reputations using our method.

Globaly, the proposed approach seems to work well for the used French health forums. Its main limitations are related to the use of a pre-established list of positive and negative expressions and to the simplicity the conducted evaluations.

4.4 Conclusion

In this chapter, we presented a method that uses the textual content of replying posts in order to infer user reputation in online forums. This method can be applied to forums where collaborative ratings are not used in order to have a similar estimation of implicit trust expressions. The proposed method consists of three main steps. First, an interaction network of "reply to" relations is extracted using both the structure of the forum and the textual content of the posts. Then, the content of each reply is evaluated in order to infer the trust or the distrust expressed in it. Finally, a reputation value is computed for each user based on the received replies. The proposed reputation metric considers propagation aspects by weighting each reply by the reputation value of its author. Manual annotations were performed in order to evaluate the proposed methods. The obtained results are promising regarding the possible application of the method.

Many perspectives can be considered in order to improve the work and to better explore the idea. First, the user's reputation can be computed for each thread or each topic. In addition to the global reputation in the whole forum, it may be interesting to compute many reputation values (one in each topic). This topic-dependent reputation can be computed by considering only replies that each user receives in the corresponding topic. Additionally, the prediction of trust, distrust and neutral posts may be improved. Crowd-sourcing tools may be used to annotate a large number of forum posts. This large corpus of annotated data will allow the learning of supervised text classification models and will enhance the evaluation (for example evaluating the systems performance on all forum posts instead of a small subset). Finally, forum moderators may be associated to define a more elaborated evaluation process of the computed reputations. This evaluation process will allow us to better asses the proposed method and the computed reputations.

In the first part of this thesis, we presented two contributions answering the question *« Who is talking? »*. We proposed two possible ways of assessing the trustworthiness of online forum users. The proposed methods are mainly based on the analyses of the textual content of the posts. Expressions of agreement, disagreement and thanking have been used in a simple way. In the next part, a deeper study of the expressed opinions inside textual contents will be presented. We will be interested in the question *« How? »*. Are the users talking positively or negatively about each other? about an object?, etc. Are they expressing fear? anger?, etc. This is the area of sentiment analysis which allows the evaluation of affect states expressed in textual messages.

Part II

Sentiment Analysis

Chapter

5 The FEEL Lexicon

Contents

5.1 Introduction 68

5.2 Compilation Process 68

5.2.1 Automatic Creation 69

5.2.2 Validating the Translations 70

5.2.3 Evaluating the Sentiments 71

5.3 Comparative Study 74

5.3.1 Evaluation in a Polarity Classification Task 76

5.3.2 Evaluation in an Emotion Classification Task 78

5.4 Conclusion 80

5.1 Introduction

Sentiment analysis allows the semantic evaluation of a piece of text according to the expressed sentiments. These sentiments are mainly conveyed by words. Therefore, sentiment lexicons play a central role in sentiment analysis. They organize lists of words, phrases or idioms into predefined classes (polarities, emotions, etc.). While considerable attention has been given to the polarity (positive, negative) of English words, only few studies were interested in the conveyed emotions (joy, anger, surprise, sadness, etc.) especially in other languages.

The sentiment evoked by a sentence is usually not the simple sum of sentiments conveyed by the words in it. However, lexicons can be very useful for sophisticated sentiment detection (either unsupervised or supervised). For example, recent studies suggest to compute sentiment based features that enhance the performance of text classification models (Hamdan et al., 2015). Indeed, it has been proved that the use of adapted sentiment lexicons can significantly improve the classification performances of text documents according to the expressed sentiments and opinions (Mohammad et al., 2015b).

To date, most existing sentiment lexicons have been created for English and for polarity (see section 2.4.1, page 26). Due to the lack of adapted French emotion lexicons, we compile a new one following the Ekman typology (Ekman, 1992). This lexicon, that we call FEEL: French Expanded Emotion Lexicon, is publicly available on the internet[1]. It has been created semi-automatically by translating and expanding to synomyms the English resource NRC-EmoLex (Mohammad and Turney, 2013). Online translators have been used at first place. Then, a human professional translator validated the obtained translations, synonyms and associated sentiments and emotions. The sentiments associated with a subset of FEEL terms have been re-evaluated by three different annotators. Finally, extensive experiments have been conducted to compare FEEL with other existing French lexicons on various French benchmarks for polarity and emotion classification (see section 2.5, page 35).

The remainder of this chapter is organized as follows. Section 5.2 describes the compilation process. Section 5.3 presents the evaluation of FEEL on a polarity and emotion classification tasks in comparison to existing lexicons. Finally, section 5.4 concludes and presents the main perspectives.

5.2 Compilation Process

The compilation process of FEEL is summarized in Figure 5.1. First, online translators have been queried automatically in order to translate and expand NRC-EmoLex terms. Then, a professional human translator checked manually all the automatically obtained entries. Finally, three human annotators evaluated the sentiments associated with a subset of FEEL terms.

[1] www.lirmm.fr/∼abdaoui/FEEL.html

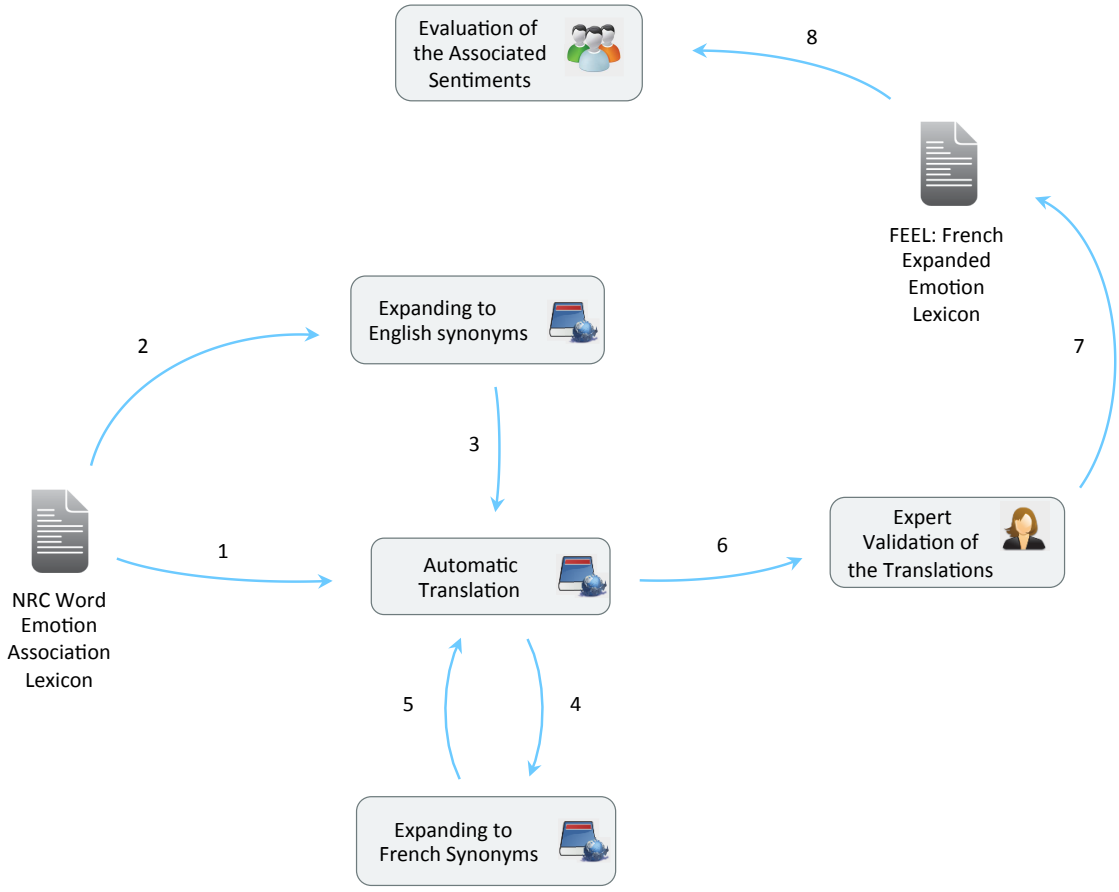


Figure 5.1: The FEEL compilation process.

5.2.1 Automatic Creation

After manually correcting some inconsistencies in NRC-EmoLex (words associated with all emotions and words associated with contradictory polarities), our aim was to automatically translate to French all of its English terms (14,182 terms). Automatic translation methods can be based on aligned resources (Och and Ney, 2004), comparable corpus (Sadat et al., 2003), or multilingual encyclopedia (Erdmann et al., 2009), etc. Since we do not have aligned resources nor comparable corpora in which we could find all the entries of the initial lexicon, we chose a different approach and used the wealth of automatic translators available online. For each entry of NRC-EmoLex, we automatically queried six online translators: Google[2], Bing[3], Collins[4], Reverso[5], Bab.la[6] and Word Reference[7]. Each English term may generate many

[2]www.translate.google.fr
[3]www.bing.com/translator
[4]www.collinsdictionary.com
[5]www.reverso.net
[6]fr.bab.la/dictionnaire
[7]www.wordreference.com

French translations. The entries that have been obtained by at least three translators have been considered pre-validated.

In order to expand our resource, we included English and French Synonyms. Synonymy corresponds to a similarity in meaning between words or phrases in the same language. Therefore, synonyms should have the same emotion and polarity class. Antonyms have not been considered since our emotion model do not support contrary emotions. In the literature, synonymy has been used to build sentiment resources by expanding seed words for which the polarity or the emotional class is already known (Strapparava et al., 2004). Here, we adopted a similar approach to expand both the English entries and the French translations. For all English entries of the original resource, we searched for synonyms using eight online websites: Reverso, Bab.la, Atlas[8], Thesaurus[9], Ortolang[10], SensAgent[11], The Free Dictionary[12] and the Synonym website[13]. The obtained English synonyms have been translated as previously described. Similarly, for all French entries, we searched for synonyms using two online websites: Ortolang and Synonymo[14]. Entries associated with contradictory polarities have been automatically removed. Finally, the automatically compiled resource contained 141,428 French entries (56,599 considered as pre-validated and 84,829 considered as non pre-validated entries).

5.2.2 Validating the Translations

In order to obtain a high quality resource and to evaluate the automatic process, we hired a human professional translator. All the automatically obtained entries have been presented to her via a web interface. For each English term, she can validate or not the automatically obtained translations, manually add a new translation and change the associated polarities and emotions. Examples of sentences using the current term have been presented in order to better understand its meaning. These sentences have been generated from the Linguee website[15]. Our professional translator worked full-time for two months. She validated less than 18% of the entries that have been obtained by less than three translators (15,091 terms), against more than 94% of ones that have been found by at least three online translators (53,277 terms). This result shows that it is possible to use online translators in order to uncostly compile good quality resources.

In addition to the validated entries based on the automatic translators, our human translator manually added 10,431 new French translations based on the displayed English terms. Finally, our resource contained 81,757 French entries (lemmas

[8] dico.isc.cnrs.fr
[9] www.thesaurus.org
[10] www.cnrtl.fr/synonymie/
[11] dictionnaire.sensagent.com/synonyme/en-fr/
[12] www.thefreedictionary.com
[13] www.synonym.com
[14] www.synonymo.fr
[15] www.linguee.fr

and flexed forms), which have been lemmatized using the TreeTagger tool (Schmid, 1994). This process generated 14,127 distinct lemmatized terms consisting in 11,979 words and 2,148 compound terms. The lemmatized terms have been associated with all the emotions of their inflected forms. Terms associated with contradictory polarities have been removed (81 terms). We considered that these terms dot not convey sentiments by their own and may be positive or negative according to their context. For example, the word “to vote” may be used either in a positive context “to vote for” or in a negative one “to vote against”.

Table 5.1 shows the repartition of the final lemmatized terms between the two considered polarities and the six basic emotions, and the intersections between them. It appears that most positive entries are associated with the emotion joy. However, some positive entries are associated with the emotions surprise, fear, sadness, anger and disgust. For example, the human translator validated the word "dive" as positive but associated with the emotion fear. On the hand, most negative entries are associated with the emotions surprise, fear, sadness, anger and disgust. Nevertheless, very few negative entries are associated with the emotion joy. For example, the word "heady" is negative but has been associated with the emotion joy. We decided not to consider these associations as inconsistent since our human translator validated them. Similarly, emotions may have common terms especially negative ones. For example, the word "accuse" is associated with the emotions anger and disgust. Finally, joy is the most pure emotion since it does not have any common entry with the remaining Ekman basic emotions.

Table 5.1: The intersections between polarities and emotions in FEEL.

	Positive	Negative	Joy	Surprise	Anger	Disgust	Sadness	Fear
Positive	5,704							
Negative	0	8,423						
Joy	514	7	521					
Surprise	435	747	0	1,182				
Anger	120	1,983	0	355	2,103			
Disgust	92	1,922	0	133	889	2,014		
Sadness	133	2,381	0	291	932	837	2,514	
Fear	223	2,976	0	657	1,335	909	1,532	3,199

5.2.3 Evaluating the Sentiments

While the professional manual translations can be considered reliable, the associated sentiments and emotions may be subjective (since performed by only one annotator). In order to evaluate the quality of our resource, the sentiments and emotions

associated with a subset of FEEL terms have been evaluated manually by three new annotators. In order to compile this subset, we selected terms that are frequent in the four French benchmarks presented in section 2.4.3. Terms that appear at least 10 times in the training set and at least 10 times in the testing set of each benchmark have been selected. Figure 5.2 shows the frequency of FEEL terms in the training set of the Climate benchmark (shown in a log_{10} scale). The horizontal line ($y = 1$) corresponds to our frequency threshold ($log_{10}(10) = 1$). Finally, 120 terms have been selected which represents less than 1% of FEEL terms. However, this subset covers almost a third of FEEL terms occurrences in the presented benchmarks. Regarding the division between the two polarities, 109 terms were initially assigned to the positive polarity against 11 terms associated with the negative one. On the other hand, each emotion of the Ekman typology has only seven terms except the emotion Anger that has four terms. Most of the terms are not associated with any emotion.

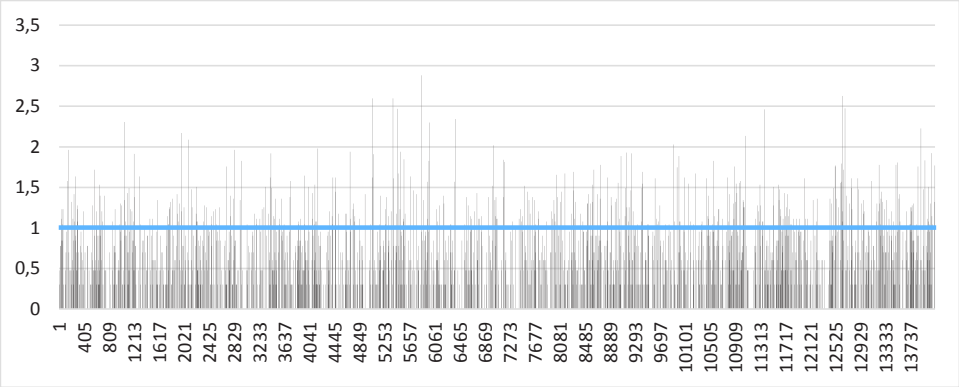


Figure 5.2: The distribution (in a log_{10} scale) of FEEL terms in the training set of the Climate benchmark.

These terms have been presented to three new annotators in order to check the associated polarities and emotions. In order to handle polysemy, two types of annotation have been performed:

- **Annotation without context:** the annotators are asked to choose the associated polarities and emotions without presenting any example to them;
- **Annotation in context:** the annotators are asked to choose the associated polarities and emotions according to the sense in the displayed sentence. Four contexts have been considered corresponding to our four benchmarks. From each benchmark, we selected the first sentence containing the corresponding term and present it as an example to the annotators.

Table 5.2 presents the agreement between the three annotators in each annotation type. First, Fleiss' kappa shows good polarity agreement and bad emotion agreement

in both annotation types. These results are similar to those obtained in (Mohammad and Turney, 2013) when building the original English NRC-EmoLex. However, Fleiss' kappa does not take into account the number of items per category. Since we have very unbalanced categories (much more terms associated with the category "no" than terms associated with the category "yes" for a given emotion), we also present the percentage of terms for which the three annotators have chosen the same category. Indeed, our three annotators agreed for most of the terms (more than 85% in each task and annotation type). Finally, our annotators suggested to include the polarity "neutral" in our future work.

Table 5.2: The annotators agreement for polarity and emotions (arithmetic mean) in each annotation type. We present the Fleiss' Kappa and the percentage of terms for which all annotators chose the same sentiment class.

	Fleiss' Kappa		**Percentage of terms with perfect agreement**	
	No context	*In context*	*No context*	*In context*
Polarity (+/−)	**0.68**	0.56	92.5%	85.4%
Emotions (yes/no)	0.22	0.18	95.4%	**95.6%**

Finally, the annotations without context have been used to evaluate the initial sentiments and emotions. A majority vote has been considered in order to extract the reference annotations. Table 5.3 presents the weighted average precisions, recalls and F1-measures for polarity and emotions. Weighted averaging is used to deal with unbalanced data sets. In our case, we used the label-frequency-based weighted average-averaging. It weighs each class results with its proportion of documents in the test set. The emotions evaluation metrics have been averaged by arithmetic mean between the six emotions. The presented results show very high consistency between the initial sentiments and those selected by at least two new annotators (majority vote).

Table 5.3: The evaluation of the sentiments associated with the chosen subset of terms.

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Polarity (+/−)	0.99	0.99	0.99
Emotions (yes/no) - mean	0.96	0.99	0.98

5.3 Comparative Study

Here we compare FEEL to the existing French sentiment lexicons presented in section 2.1. Among the four listed lexicons, only CASOAR has not been included here since it is not publicly available. The remaining three lexicons have been used in our evaluations. All of them contain lemmatized terms excepting Diko. The expressions of this last lexicon have been cleaned and grouped into lemmatized terms. Figure 5.3 presents the percentage of terms in each lexicon according to their number of words. It appears that almost all Affects and Polarimots terms are composed of only one word (100% for Polarimots and more than 99% for Affects). Then, more than 85% of FEEL terms are words and almost 15% are compound terms. Among the compound terms, 9% are composed of two words and 5% are composed of three words. Finally, only 33% of Diko terms are words. The rest are divided as follows: 31% are composed of two words, 22% are composed of three words, 8% are composed of four words, 3% are composed of five words and the remaining 3% are composed of more than five words.

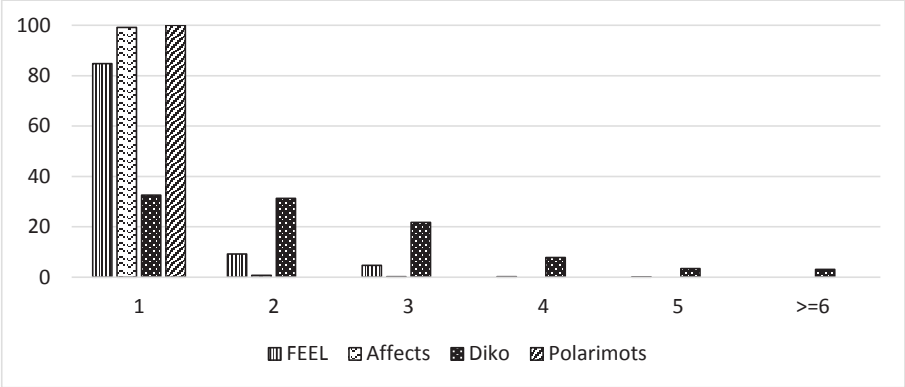


Figure 5.3: The percentage of terms in each lexicon according to their length (number of words).

Table 5.4 presents the number of terms in each lexicon and the number of common terms between each couple of lexicons. Diko is the largest resource with 382,817 lemmatized French entries. FEEL is the second largest with 14,127 terms. Polarimots and Affects lexicon contain 7,483 and 1,348 terms respectively. Diko covers almost 97% of FEEL terms (13,681 out of 14,127), almost 88% (1,182 out of 1,348) of Affects terms and more than 98% of Polarimots terms (7,359 out of 7,483). Therefore, Diko is clearly the most extensive resource but we do not have information about the proportion of noisy terms that it may contains (non-affective terms).

Table 5.5 shows the number of positive, negative and neutral terms in each lexicon. FEEL is the only lexicon that do not consider the neutral polarity. We notice that all lexicons have more negative terms than positive ones except Diko. The algorithm used for selecting the candidate terms may explain this observation (Lafourcade et al., 2015b).

Table 5.4: The intersections between the terms in each couple of lexicons.

	FEEL	Affects	Diko	Polarimots
FEEL	14,127			
Affects	559	1,348		
Diko	13,681	1,182	382,486	
Polarimots	2,747	237	7,359	7,483

Table 5.5: The number of positive, negative and neutral terms in each lexicon.

	FEEL	Affects	Diko	Polarimots
Positive	5,704	437	224,832	1,315
Negative	8,423	790	55,593	1,464
Neutral	0	121	102,061	4,704

Regarding the agreement between each couple of lexicons about the associated polarities, Table 5.6 presents the percentage of common terms having the same polarity. Neutral terms have not been considered in these calculations. Table 5.6 shows that for all couples of lexicons, more than 80% of their common positive and negative terms are associated with the same polarity. The highest agreement is observed between Diko and Polarimots with 91% of common terms associated with the same polarity.

Table 5.6: The percentage of common terms having the same polarity between each couple of lexicons.

	FEEL	Affects	Diko
Affects	89%		
Diko	83%	89%	
Polarimots	80%	86%	**91%**

Finally, all the used lexicons consider the polarity of French terms but only three give the exact emotion class (Polarimot do not consider emotions). Each one of the remaining lexicons follows its own emotional typology (FEEL: 6 emotions, Affects Lexicon: 45 emotions, Diko: more than 1,200 emotion terms).

5.3.1 Evaluation in a Polarity Classification Task

Here, we evaluate the classification gain when using features extracted from different lexicons compared to bag of word classifiers. Only positive and negative text documents of the benchmarks presented in section 2.4.3 have been used in these evaluations.

First, Support Vector Machines (SVM) have been trained on each benchmark with the Sequential Minimal Optimization method (Platt, 1999). The Weka data-mining tool (Hall et al., 2009) has been used to train these classifiers with default settings on lemmatized and lowercased text documents. A feature selection step has been performed using the Information Gain filter (words having positive Information Gain have been selected). In our experiments, we call this configuration Bag-Of-Words. Then we add to this configuration, two features from each lexicon. Indeed, we compute the number of positive words and the number of negative words according to each lexicon. These two features have been added before applying the Information Gain filter. Six other configurations have been evaluated for each benchmark corresponding to the four tested lexicons and the two additional FEEL variations: FEEL with replacement of the 120 terms from the annotation without context (FEEL-WiCxt) and in the corresponding context (FEEL-InCxt). The macro and weighted average precisions, recalls and F1-measures of these configurations applied on each corpus are presented in Tables 5.7, 5.8, 5.9 and 5.10.

Table 5.7: The polarity classification results on the benchmark See & Read.

	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Bag-Of-Words	83.5	74.2	77.4	86.2	86.9	85.8
Bag-Of-Words + FEEL	**84.5**	**76.6**	**79.5**	**87.2**	**87.8**	**87.0**
Bag-Of-Words + FEEL-WiCxt	**84.5**	**76.6**	**79.5**	**87.2**	**87.8**	**87.0**
Bag-Of-Words + FEEL-InCxt	**84.5**	**76.6**	**79.5**	**87.2**	**87.8**	**87.0**
Bag-Of-Words + Affects	84.2	75.0	78.3	86.7	87.3	86.3
Bag-Of-Words + Diko	84.0	75.6	78.7	86.8	87.4	86.5
Bag-Of-Words + Polarimots	83.5	74.2	77.4	86.2	86.9	85.8

The Bag-Of-Words configuration with lemmatization, lowercasing and especially feature subset selection represents a highly efficient baseline. Indeed, this configuration obtained high macro and weighted average precisions, recalls and F-measures on all benchmarks. Moreover, the Information Gain filter selected between 63 and 390 lemmatized words for every benchmark. Therefore, it is difficult to observe a significant gain only by adding two new features. Still, the performance gain is noticeable in all benchmarks. Almost all the lexicons induce a gain that varies from 0.1% to 7.1% in the considered evaluation metrics. If the use of lexicons obtains a little gain on the three first benchmarks (See & Read, Political Debate and Videos

Table 5.8: The polarity classification results on Political Debate.

	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Bag-Of-Words	70.2	**70.2**	70.0	70.6	70.8	70.7
Bag-Of-Words + FEEL	**70.6**	**70.2**	**70.3**	**71.0**	**71.1**	**71.0**
Bag-Of-Words + FEEL-WiCxt	70.5	70.1	70.1	70.9	**71.1**	70.9
Bag-Of-Words + FEEL-InCxt	70.4	70	70.2	70.8	71	70.8
Bag-Of-Words + Affects	70.4	70.0	70.2	70.8	71.0	70.9
Bag-Of-Words + Diko	70.4	70.0	70.1	70.8	71.0	70.8
Bag-Of-Words + Polarimots	70.2	69.9	70.0	70.6	70.8	70.7

Table 5.9: The polarity classification results on the benchmark Videos Games.

	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Bag-Of-Words	93.6	93.4	93.5	94	94	94
Bag-Of-Words + FEEL	93.5	93.5	93.5	94	94	94
Bag-Of-Words + FEEL-WiCxt	93.5	93.5	93.5	94	94	94
Bag-Of-Words + FEEL-InCxt	93.5	93.5	93.5	94	94	94
Bag-Of-Words + Affects	93.5	93.5	93.5	94	94	94
Bag-Of-Words + Diko	93.8	93.7	93.8	94.2	94.2	94.2
Bag-Of-Words + Polarimots	**94.0**	**94.0**	**94.0**	**94.4**	**94.4**	**94.4**

Games), their use induce a 7% gain on the fourth benchmark (Climate). This observation may be related to the text nature, since the fourth benchmark is the only one that contains tweets. Indeed, tweets are very short text documents (less than 140 characters) while product reviews or debate reports can contain hundreds of words. Regarding the performance of each lexicon, we notice that it depends on the benchmark. There is no lexicon that obtains the best results in all the used benchmarks. FEEL obtains the best results on two benchmarks (online reviews and debate transcriptions), Polarimots obtains the best results on Video Games and Diko on tweets. Globally, FEEL obtains very competitive results being the best on two benchmarks and second on a third one (Climate). The difference between FEEL and the best configuration is always less than 1%. Regarding the two derivations of FEEL from the re-annotation, we observe a small change in the results in comparison with the original resource. This observation may be explained by the very high consistency between FEEL-WiCxt and FEEL as presented in table 6. On the other hand, the choice of the example sentence in the annotation with a context may be unrepresentative of the term use in the whole benchmark.

Table 5.10: The polarity classification results on the benchmark Climate Polarity.

	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Bag-Of-Words	72.8	69.1	69.2	72.4	71.6	70.3
Bag-Of-Words + FEEL	76.1	74.8	75.1	76.1	76.1	75.8
Bag-Of-Words + FEEL-WiCxt	76.4	75.6	75.8	76.5	76.6	76.4
Bag-Of-Words + FEEL-InCxt	76.4	75.6	75.8	76.5	76.6	76.4
Bag-Of-Words + Affects	73.3	72.4	70.2	73.3	72.4	71.3
Bag-Of-Words + Diko	**77.8**	**76.0**	**76.4**	**77.6**	**77.4**	**77.1**
Bag-Of-Words + Polarimots	74.2	70.7	71.0	73.7	73.0	72.0

5.3.2 Evaluation in an Emotion Classification Task

Only the Climate benchmark provides emotion classes for its text documents (tweets). It uses an emotional typology divided into 18 classes as presented in Figure 2.4 page 35. As mentioned before, these emotional classes are very unbalanced. For example, only six tweets are associated with the emotion Boredom, while 2,148 tweets are labeled with the emotion Valorization. Therefore, macro averaging is not adapted in this case. Here, we only consider the weighted averaging (the label-frequency-based averaging). Regarding the lexicons, Polarimots is the only resource that do not consider emotions. We perform our evaluations using the remaining lexicons. FEEL proposes six emotion classes, Affects has 45 emotions and Diko associates its terms with 1,198 emotion expressions. We use the same baseline as in the polarity classification task (Bag-Of-Words). To this configuration, we evaluate the add of features extracted from each emotion lexicon. These features represent the number of terms expressing each emotion. Therefore, six features are added for FEEL, FEEL-WiCxt and FEEL-InCxt, 45 features are added for Affects and 1,198 features are added for Diko. The feature selection step is applied after adding these features. Lemmatization and lowercasing are also performed when searching the emotion terms inside the tweets. Table 5.11 presents the emotion classification results when considering the 18 original emotion classes.

It appears that all emotion lexicons improve significantly the classification results. The gain is between 5.7% and 12.9% in weighted average precision, between 3.9% and 5.3% in weighted average recall and between 5% and 7.1% in weighted average F-measure. Diko obtains the highest weighted average recall but the lowest weighted average precision (due to its large number of entries). FEEL is ranked third but close to the best configuration for each evaluation metric. FEEL-WiCxt and FEEL-InCxt improve slightly the classification results. However, the emotional typology of the Climate corpus (18 classes) do not refer to a well-known classification. We are evaluating FEEL on classes that it does not consider. In order to have an estimation of each lexicon performance according to the Ekman emotional classes, we perform the same experiments but when considering only the four Ek-

Table 5.11: The emotion classification results when considering 18 emotional classes.

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Bag-Of-Words	46.9	49.7	39.7
Bag-Of-Words + FEEL	50.8	53.6	44.7
Bag-Of-Words + FEEL-WiCxt	51.1	53.9	45.1
Bag-Of-Words + FEEL-InCxt	50.9	53.7	45
Bag-Of-Words + Affects	50.9	53.8	45.4
Bag-Of-Words + Diko	**52.6**	**55.0**	**46.8**

man emotions that are present is the Climate corpus. The division of the considered tweets between the emotions (surprise, anger, fear and sadness) are presented in Figure 2.4. In addition to the Bag-Of-Words configuration, we evaluate the add of six features for FEEL, FEEL-WiCxt and FEEL-InCxt, 45 features for Affects and 1,198 features Diko.

Table 5.12: The emotion classification results when considering 4 emotional classes.

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$
Bag-Of-Words	74	70	68.2
Bag-Of-Words + FEEL	**74.3**	**74.4**	**72.8**
Bag-Of-Words + FEEL-WiCxt	73.6	73.5	72.2
Bag-Of-Words + FEEL-InCxt	73.6	73.5	72.2
Bag-Of-Words + Affects	69.1	69.5	69.2
Bag-Of-Words + Diko	71.7	68.6	66

Table 5.12 shows that FEEL obtained the best results. It generates a gain of 0.3% in weighted average precision, 4.4% in weighted average recall and 4.6% weighted average F1-measure in comparison to the Bag-Of-Words configuration. FEEL-WiCxt and FEEL-InCxt come second with close precisions, recalls and F1-measures. Finally, Affects and Diko generate a decrease in the evaluation metrics, which suggests that these lexicons are not adapted to the Ekman emotions. Since Affects and Diko propose a finer emotional typology, we may think that this should not influence the classification performance with less emotional classes. Even though, FEEL significantly outperforms these two lexicons for the available Ekman emotions (four out of six). Since Climate is the only available French benchmark for emotion classification, we could not test FEEL on the Ekman emotions: joy and disgust.

5.4 Conclusion

In this chapter, we presented the elaboration and the evaluation of a new French sentiment lexicon. This lexicon considers both polarity and emotion following the Ekman typology. It has been compiled semi-automatically by translating and expanding to synonyms the English lexicon NRC-EmoLex (Mohammad and Turney, 2013). A human professional translator supervised all the automatically obtained terms and enriched them with new manual terms. She validated more than 94% of the entries that have been found by at least three online translators, and less than 18% of the ones that have been obtained by less than three translators. This result shows that online translators can be used to inexpensively compile such resources using appropriate heuristics and thresholds. The obtained lexicon (FEEL) contains 14,127 French entries where around 85% are single words and 15% are compound words. It has been made publicly available for the sentiment analysis community.

While the professional manual translations can be considered reliable, the associated sentiments and emotions may be subjective. Therefore, three new annotators re-evaluated the polarities and emotions associated with a subset of 120 terms. This step showed high consistency between the initial sentiments and the new ones. Then, we performed exhaustive evaluations on French benchmarks for polarity and emotion classification. We compared our results with those obtained by existing French sentiment lexicons. In order to represent each lexicon we used the number of terms expressing each sentiment as features. The obtained results highlight that FEEL improves the classification performances on various benchmarks dealing with very different topics. Indeed, FEEL obtained competitive results for polarity (being first on two benchmarks and always very close to the best configuration) and the best results for emotion (when considering the available Ekman emotional typology). It could be noticed that the classification gain is more important for short text documents such as tweets.

As perspective, it would be interesting to compile a benchmark of annotated French text documents according to the six basic Ekman emotions. This benchmark will allow us to evaluate our lexicon using the same emotional typology. Similar benchmarks have been compiled for English (Strapparava et al., 2004) following the Ekman basic classes but not for French. On the one hand, crowdsourcing tools can be used to obtain large number of manual annotations. Indeed, they gather contributors from all over the world to perform dedicated tasks. Using these tools, we can obtain a large number of annotated documents and a large number of annotations per document. In order to enhance the quality of the annotations, we can upload an annotated gold dataset and remove the annotations made by contributors having very bad accuracy based on the gold dataset. On the other hand, we can scroll the Twitter API with the names of the emotions as hashtags (#joy, #surprise, #anger, #sadness, #fear and #disgust). Indeed, (Mohammad and Kiritchenko, 2015) used this process for English tweets and obtained a benchmark of a good quality.

Using FEEL, we built French sentiment classification systems that participated to the evaluation campaign DEFT'15. Among 22 teams that have registered to the challenge, our systems were ranked first in subjectivity classification, third in polarity classification and fifth in emotion classification (when considering 18 classes). The proposed systems are also based on SVM classifiers with more elaborated features and tuning methods. In the next chapter, we will present the features and methods and resources used in these sentiment classification systems. Extensive experiments have been conducted on the French sentiment benchmarks of DEFT'07 and DEFT'15. A feature engineering process has been applied to detect the best features and methods for each benchmark. The FEEL lexicon has been used in these systems as well as the remaining French sentiment lexicons.

Chapter

6 French Sentiment Classification

Contents

6.1 Introduction 84
6.2 Features and Methods 84
6.2.1 Word Ngrams 85
6.2.2 Preprocessings 85
6.2.3 Handling Negation 85
6.2.4 Lexicon Features 85
6.2.5 Syntactic Features 86
6.2.6 Word Embeddings 86
6.2.7 Feature Subset Selection 86
6.2.8 Classifier 86
6.3 Experimentations 87
6.3.1 Tuning on Cross Validation 87
6.3.2 Evaluating the Selected Configurations 90
6.4 Conclusion 94

6.1 Introduction

Due to its large number of applications, sentiment analysis has received much attention from both scientific and economic communities in the last decade. Most studies concerned sentiment classification which represents an essential task in sentiment analysis. Sentiment classification consists in classifying text documents according to their polarity, subjectivity, emotion, etc. Most of the work dealt with the classification of English text documents. After compiling a specific French sentiment lexicon, our aim is to build French sentiment classification systems. In this chapter, we evaluate different combinations of features and methods for French sentiment classification. Support Vector Machines have been trained with different combinations of features (word ngrams, syntactic features, etc.), pre-processings (lemmatization, slang replacement, etc.), methods (feature subset selection, handling negation, etc.), resources (polarity and emotion lexicons) and parameters.

Extensive experiments have been conducted on the French sentiment benchmarks presented in section 2.4.3, page 33. Three sentiment classification tasks have been considered: polarity classification (2 and 3 classes), subjectivity classification (4 classes) and emotion classification (18 classes). In order to choose the best configuration for each benchmark, we propose a feature engineering process and apply it by cross validation on the training sets. Only features and methods that improve the classification results are selected. When applied to the test sets, the selected configurations obtained comparable results to the best systems in the DEFT'07 and DEFT'15 challenges. Our systems outperform (on two out of the six considered benchmarks) the best results obtained in these challenges. The sources code of these systems is available on GitHub[1]. One can reproduce the presented results by editing a configuration file.

The remainder of this chapter is divided as follows. Section 6.2 presents the evaluated features, pre-processings and methods for French sentiment classification. Section 6.3.1 describes a feature engineering process in order to chose the best features and tune the classification models by cross validation. Section **??** presents and discusses the obtained results. Finally, section 6.4 concludes and give our main prospects.

6.2 Features and Methods

The following features and methods have been implemented and evaluated for French sentiment classification.

[1] github.com/amineabdaoui/SentimentClassification

6.2.1 Word Ngrams

Word Ngrams are considered to be the basic features in text classification including sentiment classification. It has been reported that the use of binary representation works better than frequency-based representations for sentiment classification (Pang et al., 2002, Liu, 2012). Therefore, we consider the presence or absence of unigrams, bigrams and both unigrams and bigrams.

6.2.2 Preprocessings

As mentioned in (Haddi et al., 2013), texts from Social Media have some linguistic peculiarities that may affect the sentiment classification performance. For this reason, the following pre-processings are applied. Some of them have been used in chapter 3 in the prediction of the user expertise (see section 3.2.2, pages 42).

1. Hyperlinks, emails and pseudonyms normalization;
2. Slang replacement with the corresponding text using a pre-established list;
3. Lower-casing;
4. Lemmatization using TreeTagger (Schmid, 1994);
5. Stop words removal.

6.2.3 Handling Negation

Words appearing within the scope of a negation term receive the suffix "_neg". As in (Pang et al., 2002), we assume that the scope begins with the negation term and ends with a punctuation mark. This method allows the classification model to distinguish between words used in positive and negative context. It has been wildly applied to handle negation in English sentiment classification (Mohammad et al., 2013, Hamdan et al., 2015). Here, we evaluate its use in French sentiment classification. For more information about this method, please visit Christopher Potts' sentiment tutorial[2].

6.2.4 Lexicon Features

Number of words expressing each sentiment class (polarity or emotion) according to a given lexicon (FEEL-pol: 2 features, FEEL-emo: 6 features, Affects-pol: 3 features, Affects-emo: 45 features, Diko-pol: 3 features, Diko-emo: 1,198 features and Polarimots-pol: 3 features).

[2]sentiment.christopherpotts.net/lingstruc.html#negation

6.2.5 Syntactic Features

The syntactic features presented in (Mohammad et al., 2013) have been implemented and tested:

1. Elongated words: number of words containing repeated characters (more than three identical consecutive characters);
2. Punctuation: presence or absence of an exclamation point or a question mark;
3. Capitalization: number of words with all characters in upper case;
4. Smileys: presence or absence of positive and negative smileys;
5. Hashtags: number of hashtags;
6. Negation: number of negation terms;
7. POS tags: presence or absence of each part of speech tag.

6.2.6 Word Embeddings

We evaluate the use of word embeddings learned in (Rouvier et al., 2015). The authors collected 16 millions French tweets using sentiment keywords (good, like, etc.) and smileys (;), :-), etc.). Then, Word2Vec has been used to learn these word embeddings using the Continuous Bag of Words approach (Mikolov et al., 2013a). The embedding vector size has been set to 100, which means that each word has been represented in a space of 100 dimensions. In order to represent our textual documents (which have an unfixed number of words), we evaluate the use the min, max and average convolutional layers described in (Collobert et al., 2011).

6.2.7 Feature Subset Selection

Since the number of word ngrams depends on the size of the training data, the dimensionality of the features may grow dramatically introducing many redundant and irrelevant features. Therefore, a feature subset selection step has been tested. The Information Gain (IG) method (Mitchell, 1997) has been used in order to rank the features according to their predictive power. Features having positive IG have been selected for each benchmark.

6.2.8 Classifier

The chosen classification model is SVM with the Sequential Minimal Optimization method (Platt, 1999). This algorithm appeared to be effective in text categorization and especially sentiment classification (Mohammad et al., 2013). Furthermore, it remains robust on large feature spaces. We used the Weka Data Mining tool (Hall et al., 2009) to learn this classification model in each experiment with a polynomial kernel. The complexity parameter (C) has been estimated on cross validation.

6.3 Experimentations

In this section, we present the conducted experiments on the considered French sentiment benchmarks. First, we search for the best configuration of features that work best for each benchmark by cross validations on the training sets. Then, we evaluate the selected configurations and compare the our results with those obtained at each benchmark.

6.3.1 Tuning on Cross Validation

In order to find the best configuration of features, methods and parameters for each benchmark, k-fold cross validations have been performed on the training sets. 10-folds cross validations have been applied for all benchmarks except the Climate - Emotion benchmark. This last contains many classes with less than 10 tweets (classes with $log_{10} < 1$ in figure 2.4, page 35). Therefore, 3-fold cross validation has been applied to this benchmark. Our feature engineering process has been divided into 8 steps as shown in figure 6.1. The characteristics of each step have been tested independently and only those that improve the results according to the chosen evaluation metric have been selected. The details about characteristics tested at each step have been described in the previous section.

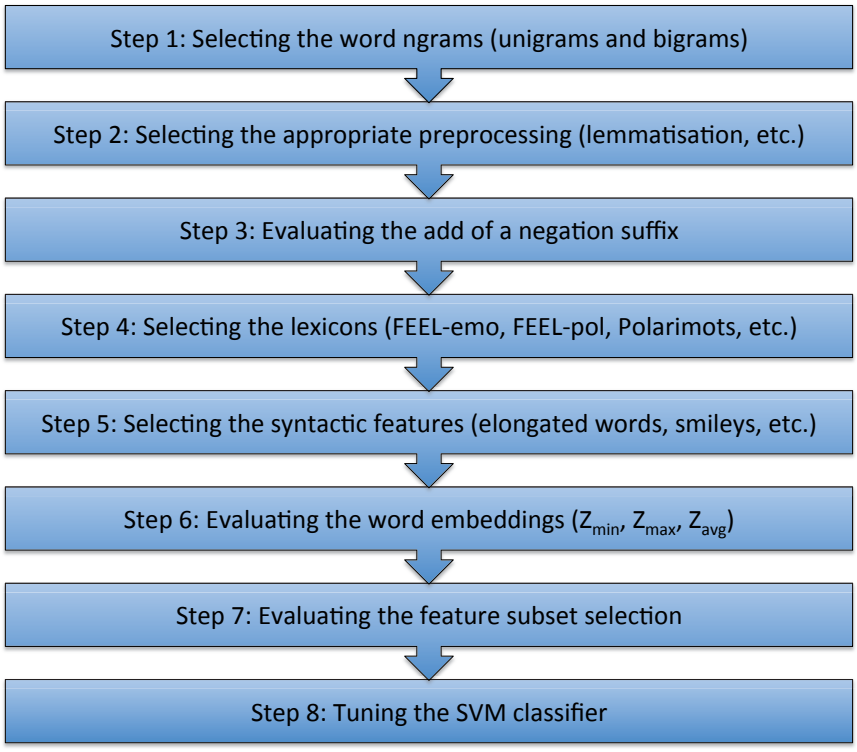


Figure 6.1: Steps of our Feature Engineering Process.

At each step, only pre-processings, features or parameters that improve the results have been selected (if any). The best configuration at step n is used as baseline in testing the features and parameters of the following step $n + 1$. The considered evaluation metric to make the selection is the weighted average averaged F1-measure. However, since macro averaged precision has been selected in DEFT'15 and since feature selection is known to improve the precision but not the recall, we considered macro precision for DEFT'15 benchmarks (polarity, subjectivity and emotion classification) starting at the feature selection step.

Table 6.1 presents the selected features and parameters for each benchmark. It appears that both unigrams and bigrams have been selected for DEFT'07 benchmarks, but only unigrams have been selected for DEFT'15 benchmarks. This observation may be explained by the very different nature of their documents (reviews and debates vs tweets). DEFT'07 benchmarks are characterized by their length (hundreds of words per document), while DEFT'15 benchmarks are characterized by their shortness (less than 140 characters). The contrast between these two categories of text documents can also be observed throw the pre-processing step. For example, hyperlinks, emails and pseudonyms normalizations have been selected for DEFT'15 benchmarks on the contrary of DEFT'07 benchmarks. Indeed, the Climate benchmarks contain many hyperlinks that often point to newspaper websites and pseudonym tags when replying or re-tweeting other users statements. Slang replacement has only been selected for the See & Read benchmark, since very few slang expression can be observed on the other benchmarks. Finally, lower-casing has been selected for all benchmarks. Regarding the remaining steps, lexicon based and syntactic based features are more useful in the classification of tweets. Most of them have been select for DEFT'15 benchmarks, while very few have been selected for DEFT'07 benchmarks. These observations can be very helpful to quickly choose the best features and pre-processings for French sentiment classification according to the text length and nature.

In order to present the effect of each category of features/methods on the results, tables 6.2 – 6.7 show the weighted average and macro precisions, recalls and F1-measures at the end of each step of our process. The numbers between brackets represent the difference between the corresponding step and the previous one for each evaluation metric. If no feature has been selected at a given step, the presented results are equal to those obtained at the end of the previous step, and the difference between the two steps is equal to 0.

For all benchmarks, ngrams have an important impact (step 1). The results obtained at the end of this first step are close to those obtained at the end of the whole process, especially on the Videos Games benchmark (the improve after the 7 following steps does not exceed 0.4% in terms of weighted average F1-measure). The pre-processings (step 2) improve the results for all benchmarks. This improve is higher for See & Read and the Climate benchmarks (product reviews and tweets). Handling negation by adding a "_neg" suffix (step 3) seems to have a small impact on the results (do not exceed 0.1%). The same observation has been observed in

Table 6.1: The selected features and parameters by cross validation on the training set of each benchmark.

		DEFT'07			**DEFT'15**		
		See & Read	Video Games	Parliamentary Debate	Climate Polarity	Climate Subjectivity	Climate Emotion
Step 1	Unigrams	√	√	√	√	√	√
	Bigrams	√	√	√			
Step 2	Hyperlinks				√	√	√
	Emails				√	√	
	Pseudonyms				√		
	Slangs	√					
	Lemmatization	√			√	√	√
	Lower-casing	√	√	√	√	√	√
	Stop words	√			√		
Step 3	Negation		√	√			√
Step 4	FEEL-pol	√	√		√	√	√
	FEEL-emo				√	√	√
	Affects-pol	√	√		√	√	
	Affects-emo				√	√	√
	Diko-pol		√		√	√	
	Diko-emo						√
	Polarimots		√		√	√	√
Step 5	Capitalization				√	√	√
	Elongated words				√		
	Hashtags				√	√	
	Negation terms				√		
	Punctuation		√	√		√	√
	POS tags						
	Smileys				√	√	
Step 6	Word embeddings Z_max			√			
	Word embeddings Z_min						
	Word embeddings Z_avg					√	√
Step 7	Feature selection			√	√	√	
Step 8	Complexity parameter	1	1	0.05	0.11	0.08	0.2

(Vincent and Winterstein, 2013) when applying this negation handling method for French text documents. Curiously, Pang's negation handling method performs well for English but not for French. Lexicon features (step 4) improve the classification results of DEFT'15 benchmarks (between 1.9% and 2.8% in weighted average F1-measure), but not those of DEFT'07. This finding suggests that lexicon based features have a higher effect on short French texts, which joins the claims of (Hamdan et al., 2015) for short English texts. This phenomenon may be due to the considered features which are based on the count of sentiment terms. Once again, these observations tend to confirm the contrast between long texts and short ones (tweets). The word Embeddings (step 6) have a small influence on the results (we notice a small improve in macro precision for DEFT'15 that do not exceed 0.2%). This observation may be due to the used representation (min, max and avg). The feature selection (step 7) has a high influence on the benchmarks Parliamentary Debate (+2.1% in weighted average F1-measure), Climate - Polarity (+2.9% in macro precision) and Climate - Subjectivity (+3.7% in macro precision) but has not been considered on the remaining benchmarks. As in (Vincent and Winterstein, 2013) we highlight the effect of feature subset selection in French sentiment classification. Finally, the complexity parameter estimation (step 8) significantly improves the results on the benchmarks Parliamentary Debate (+2.4% in weighted average F1-measure), Climate - Subjectivity (+4% in macro precision) and Climate - Emotion (+3.9% in macro precision).

Table 6.2: The obtained results after each step by 10-folds cross validation on the benchmark See & Read (3 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Step 1	62.7	63.2	62.4	60.9	55.3	57.1
Step 2	64.1 (1.4)	64.8 (1.6)	63.9 **(1.5)**	63.6 (2.7)	58.2 (2.9)	60.1 (3)
Step 3	64.1 (0)	64.8 (0)	63.9 (0)	63.6 (0)	58.2 (0)	60.1 (0)
Step 4	64.2 (0.1)	64.9 (0.1)	64 (0.1)	63.8 (0.2)	58.3 (0.1)	60.2 (0.1)
Step 5	64.2 (0)	64.9 (0)	64 (0)	63.8 (0)	58.3 (0)	60.2 (0)
Step 6	64.2 (0)	64.9 (0)	64 (0)	63.8 (0)	58.3 (0)	60.2 (0)
Step 7	64.2 (0)	64.9 (0)	64 (0)	63.8 (0)	58.3 (0)	60.2 (0)
Step 8	64.2 (0)	64.9 (0)	64 (0)	63.8 (0)	58.3 (0)	60.2 (0)

6.3.2 Evaluating the Selected Configurations

Once the appropriate configurations have been selected by cross validation, classification models can be learned on the training sets and applied to the test sets. Table 6.8 presents the obtained results in terms of weighted average and macro precision, recall and F1-measure. Our aim is to compare our results to those obtained at the above mentioned challenges. However, in each challenge the results of only one evaluation metric have been published (micro F1-measure in DEFT'07 and macro

Table 6.3: The obtained results after each step by 10-folds cross validation on the benchmark Videos Games (3 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Step 1	82.2	81.8	81.8	83.1	80.4	81.5
Step 2	82.5 (0.3)	82.1 (0.3)	82 (0.2)	83.5 (0.4)	80.9 (0.5)	81.9 (0.4)
Step 3	82.5 (0)	82.1 (0)	82.1 (0.1)	83.5 (0)	80.9 (0)	81.9 (0)
Step 4	82.6 (0.1)	82.1 (0)	82.1 (0)	83.5 (0)	81 (0.1)	82 (0.1)
Step 5	82.6 (0)	82.2 (0.1)	82.2 (0.1)	83.6 (0.1)	81.1 (0.1)	82.1 (0.1)
Step 6	82.6 (0)	82.2 (0)	82.2 (0)	83.6 (0)	81.1 (0)	82.1 (0)
Step 7	82.6 (0)	82.2 (0)	82.2 (0)	83.6 (0)	81.1 (0)	82.1 (0)
Step 8	82.6 (0)	82.2 (0)	82.2 (0)	83.6 (0)	81.1 (0)	82.1 (0)

Table 6.4: The obtained results after each step by 10-folds cross validation on the benchmark Parliamentary Debate (2 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Step 1	73.2	73.1	73.1	72	72	72
Step 2	73.7 (0.5)	73.6 (0.5)	73.6 (0.5)	72.5 (0.5)	72.5 (0.5)	72.5 (0.5)
Step 3	73.7 (0)	73.6 (0)	73.7 (0.1)	72.5 (0)	72.6 (0.1)	72.5 (0)
Step 4	73.7 (0)	73.6 (0)	73.7 (0)	72.5 (0)	72.6 (0)	72.5 (0)
Step 5	73.7 (0)	73.7 (0.1)	73.7 (0)	72.6 (0.1)	72.6 (0)	72.6 (0.1)
Step 6	73.7 (0)	73.7 (0)	73.7 (0)	72.6 (0)	72.6 (0)	72.6 (0)
Step 7	75.8 (2.1)	75.8 (2.1)	75.8 (**2.1**)	74.8 (2.2)	74.7 (2.1)	74.7 (2.1)
Step 8	78.2 (2.4)	78.3 (2.5)	78.2 (**2.4**)	77.6 (2.8)	76.9 (2.2)	77.2 (2.5)

Table 6.5: The obtained results after each step by 10-folds cross validation on the benchmark Climate - Polarity (3 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Step 1	65.1	65.1	65	65	63.2	63.8
Step 2	68.6 (3.5)	68.5 (3.4)	68.3 **(3.3)**	68.4 (3.4)	66.7 (3.5)	67.3 (3.5)
Step 3	68.6 (0)	68.5 (0)	68.3 (0)	68.4 (0)	66.7 (0)	67.3 (0)
Step 4	71.3 (2.7)	71.2 (2.7)	71.1 (**2.8**)	71.1 (2.7)	70.1 (3.4)	70.5 (3.2)
Step 5	71.3 (0)	71.3 (0.1)	71.2 (0.1)	71.2 (0.1)	70.1 (0)	70.5 (0)
Step 6	71.3 (0)	71.3 (0)	71.2 (0)	71.2 (0)	70.1 (0)	70.5 (0)
Step 7	72.2 (0.9)	71.5 (0.2)	71 (-0.2)	73.1 (**2.9**)	68.6 (-1.5)	70 (-0.5)
Step 8	73.3 (1.1)	73.1 (1.6)	72.8 (1.8)	73.6 (0.5)	71.3 (2.7)	72 (2)

precision in DEFT'15). Therefore, we compare our results according to the selected evaluation metric. The micro F1-measures obtained by our models have calculated. Figures 6.2 and 6.3 present for each benchmark our results, those of the best performing system and the averaged results of all the systems that participated to the corresponding challenge.

Table 6.6: The obtained results after each step by 10-folds cross validation on the benchmark Climate - Subjectivity (4 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Step 1	67.4	68	67.1	68.6	53.8	57.8
Step 2	70 (2.6)	70.7 (2.7)	70 **(2.9)**	65.1 (-3.5)	55.3 (1.5)	58.3 (0.5)
Step 3	70 (0)	70.7 (0)	70 (0)	65.1 (0)	55.3 (0)	58.3 (0)
Step 4	72.1 (2.1)	72.5 (1.8)	71.9 **(1.9)**	69.1 (4)	57.7 (2.4)	60.8 (2.5)
Step 5	72.2 (0.1)	72.6 (0.1)	72 (0.1)	69.3 (0.2)	57.7 (0.1)	60.8 (0)
Step 6	72.3 (0.1)	72.7 (0.1)	72.2 (0.2)	69.5 (0.2)	58 (0.3)	61.2 (0.4)
Step 7	72.9 (0.6)	72.7 (0)	71.2 (-1)	73.2 (**3.7**)	54.5 (-3.5)	59.1 (-2.1)
Step 8	72.5 (-0.4)	71.2 (-0.5)	68.9 (-1.3)	77.2 (**4**)	50.7 (-3.8)	55.6 (-3.5)

Table 6.7: The obtained results after each step by 3-folds cross validation on the benchmark Climate - Emotion (18 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Step 1	57.4	59.9	56.6	37.2	23.4	27.1
Step 2	60 (2.6)	62.6 (2.7)	60.1 **(3.5)**	37.1 (-0.1)	25.7 (2.3)	29 (1.9)
Step 3	60.1 (0.1)	62.7 (0.1)	60.2 (0.1)	37.1 (0)	25.8 (0.1)	29.1 (0.1)
Step 4	62.4 (2.3)	65 (2.3)	62.6 **(2.4)**	38.6 (1.5)	27.3 (1.5)	30.5 (1.4)
Step 5	62.7 (0.3)	65.2 (0.2)	62.8 (0.2)	38.8 (0.2)	27.4 (0.1)	30.6 (0.1)
Step 6	62.7 (0)	65.3 (0.1)	62.9 (0.1)	38.8 (0)	27.4 (0)	30.7 (0.1)
Step 7	62.7 (0)	65.3 (0)	62.9 (0)	38.8 (0)	27.4 (0)	30.7 (0)
Step 8	63.7 (1)	65.6 (0.3)	63.3 (1.4)	42.7 (**3.9**)	27.9 (0.5)	31.4 (0.7)

Table 6.8: The results obtained by the selected configurations on each benchmark.

	Weighted Avg.			**Macro**		
	P_{wa}	R_{wa}	F_{wa}	P_{ma}	R_{ma}	F_{ma}
See & Read	64.1	64.6	63.8	62.9	56.7	58.8
Videos Games	74.9	74.6	74.6	75.5	73.3	74.3
Parliamentary Debate	73.7	73.7	73.7	73.1	73.2	73.1
Climate - Polarity	71.2	69.1	67.7	72.7	64.8	66.3
Climate - Subjectivity	59.2	58.9	51.8	59.4	42.7	41.7
Climate - Emotion	57.9	61.1	58.1	31.8	24.7	26.8

Globally, it appears that our systems obtained comparable results to the best performing systems at each challenge. Regarding DEFT'07 benchmarks, our results outperform those that obtained the highest micro F1-measures on the benchmarks See & Read and Parliamentary Debate. On the benchmark Videos Games, our selected configuration obtained close micro F1-measure to the best system submitted to the challenge for this benchmark. The average micro F1-measure of the submitted systems is noticeably lower than the one obtained by selected configuration.

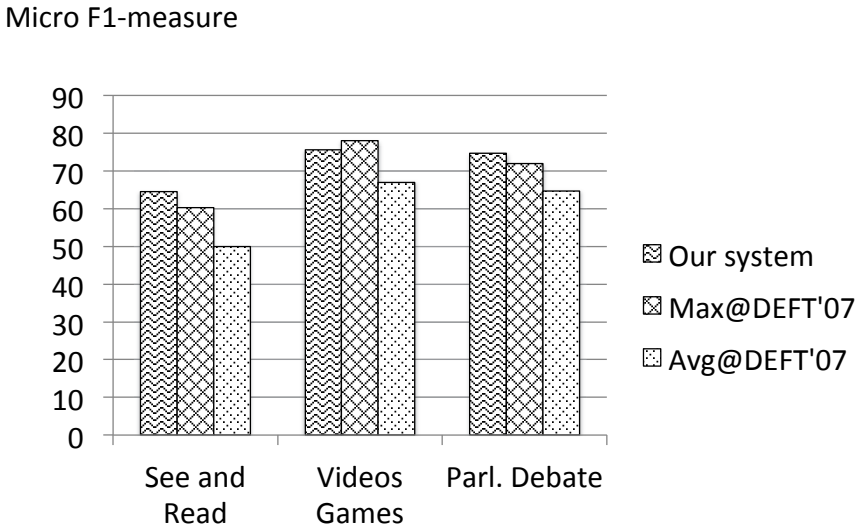


Figure 6.2: The F1-measures obtained by our system compared to the maximum and average valudes obtained at DEFT'07 for each benchmark.

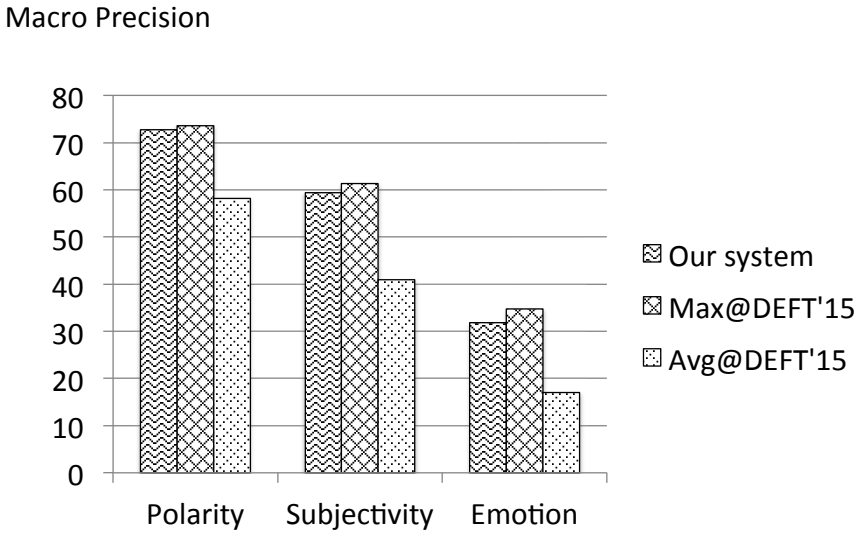


Figure 6.3: The F1-measures obtained by our system compared to the maximum and average valudes obtained at DEFT'07 for each benchmark.

On DEFT'15 benchmarks, our systems obtained close macro precisions to the best ones submitted to this challenge. The difference between our results and the average results of the systems submitted to this challenge is always greater than 10%. In addition to the findings mentioned before (best features for short and long text documents), the presented results highlight that the proposed feature engineering process can be used to build efficient sentiment classification systems.

6.4 Conclusion

In this chapter, we presented an experimental study to discover the best features and methods for French sentiment classification at the document level. Reference benchmarks from different natures (tweets, product reviews, etc.) have been used in these experiments. We implemented and evaluated a variety of pre-processings (lemmatization, slang replacement, etc.), features (syntactic features and lexicon based features) and methods (handling negation, parameter estimation, etc.). A feature engineering process has been applied by cross validation on the training sets in order to find the best configuration for each benchmark. The results of the selected configurations are comparable to the the best results obtained during French sentiment classification challenges on the same benchmarks. The source code is publicly available on GitHub. The presented results can be reproduced by editing a configuration file.

Our experiments showed that there is a clear contrast between the features and methods selected for long text documents and those selected for short ones. For example, both unigrams and bigrams are selected for long texts, while only unigrams are selected for short ones. Furthermore, lexicon based features can significantly improve the classification performances of short texts but not long ones. On the other hand, the nature of the text (formal/informal) also influences the feature choice. For example, pre-processings are very useful for informal texts (tweets and product reviews) but not on formal ones (parliamentary debates). Finally, our experiments showed that the Pang method (Pang et al., 2002) for handling negation seems to have a small impact in French sentiment classification. Curiously, this method performs well for English text classification (Mohammad et al., 2013, Hamdan et al., 2015) but not for French (Vincent and Winterstein, 2013). All these findings can be very helpful in order to quickly build efficient sentiment classification models according the text nature. Whenever it is possible, cross validation may be used to select the best features and to tune the used classifiers. For better results, many cross validations can be applied on the same training data in order to select different folds at each cross validation, but this process may consume much more computation resources.

The systems learned in this chapter have been made online on a dedicated web-platform[3]. For now, this platform allows registered users to use our models or learn their own French sentiment classification models. A demo of the tweet classification system is available without registration. It is possible to write the text directly using the web interface or to upload a file containing multiple text documents. Registered users can build their own models on the web interface to take advantage from the implemented features and methods and the used resources. In order to build new models, users have to upload tagged datasets and choose between a default mode and an advanced mode. Two default modes have been proposed (tweets and free text) with adequate configurations based on the findings discovered in this chapter. In the

[3]advanse.lirmm.fr:8081/sentiment-analysis-webpage/index

advanced mode, users have to select the different features and methods by themselves and lunch the learning process. This web-platform is still under development. The most important feature that should be added is to provide a RESTful API that will allow users to automatically query the stored models. The name of the model and the text to classify can be passed as parameters.

Other perspectives concern the used methods for sentiment classification. First, the method used to handle negation may be improved. For example, we may evaluate the use of semantic parsing to detect the scope of negators. Then, in order to use the Word Embeddings with an SVM classifier, we were forced to pass by a representation of the whole text document instead of a representation for each word. However, much information may be lost when using a document representation. Therefore, it will be interesting to evaluate the use the Word Embeddings directly using Deep Neural Networks. Furthermore, recent studies suggest that Deep Neural Networks may outperform SVM classifiers (Nakov et al., 2016). Finally, in this chapter we used already trained Word Embeddings of a fixed vector size (100 dimensions). It may be interesting to learn our own Word Embeddings to evaluate different sizes of the embedding vector. A crawler has been implemented to collect tweets with specific characteristics (language, keywords, etc.). It can be used to collect a large number of French tweets for learning our own Word Embeddings. This task may require much computational resources but the results are promising.

General Conclusion and Future Work

Chapter

7 Conclusion and Future Work

Contents

7.1 Thesis Summary 100

7.1.1 Part I: User Expertise and Reputation 100

7.1.2 Part II: Sentiment Analysis 101

7.2 Future Research 102

7.2.1 User Mining by Combining Multiple Sources 102

7.2.2 Multi-modal Social Media Mining 102

7.2.3 Enhancing Sentiment Analysis 103

7.2.4 Ethical Considerations in Social Media Mining 103

7.1 Thesis Summary

In this thesis, we presented the construction and evaluation of methods and resources for French Social Media Mining. Our contributions concerned two main parts related to (i) the expertise and reputation of the posting users and (ii) the sentiments they express in their posts. In the following, we summarize the contributions made in each part.

7.1.1 Part I: User Expertise and Reputation

First, we presented a content-based approach for predicting the medical expertise from a given forum post. The proposed idea is to use forums that hire medical experts and indicate their role, in order to learn classification models. The learned models enable to categorize expert and non-expert posts in other online discussions. Two French forums where patients and medical experts participate in the discussions have been used in order to test our method. The conducted experiments have shown that models trained on appropriate websites, where many medical experts participate in various discussions, may be used efficiently on other websites. The best results were obtained using the bag of words as features. Analyzing the misclassified posts allowed us to find out that medical experts may write posts in online health forums even if their medical role is not indicated on the website. Moreover, this analysis showed that the expertise of a user may change according to the discussed topic.

Then, a collaborative content-based method have been proposed to compute the user reputation in online forums. The basic idea consists in detecting positive and negative replies addressed to each user, and aggregate them in order to infer his/her reputation. Since the recipient of each forum message is rarely known in many online forums, we proposed a rule-based heuristic that uses both the forum structure and the messages content. The positive and negative replies have been detected using pre-established lists of agreement, disagreement and thanking expressions. Manual annotations have been conducted in order to evaluate each step of the proposed approach. The rule-based heuristic and the system detecting positive and negative replies obtained satisfactory results. In order to compute the user reputation, we proposed a measure that aggregates positive and negative replies using the constructed network of replies. The proposed measure includes propagation aspects by taking into account the reputation of the author of each positive or negative reply. Concretely, it weighs each reply by the reputation of its author. Therefore, the replies posted by good users will have more weight than those posted by bad ones. The proposed measure can be computed by successive iterations until convergence. The evaluation of the computed reputations on two French health forums showed consistent results.

7.1.2 Part II: Sentiment Analysis

The second part of this thesis concerned the development of systems and resources for French Sentiment Analysis. Due to the lack of adapted sentiment and especially emotion lexicons for French, we started by compiling a new one that we called FEEL (French Expanded Emotion Lexicon). It has been created semi-automatically by translating and expanding its well-known English counterpart NRC EmoLex. Online translators have been queried automatically to translate its English entries and expand them with their synonyms. Then, we hired a professional human translator who checked manually all the automatically obtained French entries. She validated more than 94% of the entries returned by the majority of the translators. This result highlights that the queried websites may be used efficiently to automatically translate English resources with adequate heuristics at low cost. After that, the sentiments associated with a subset of FEEL terms have been re-evaluated by three different annotators. The results showed high consistency between the new annotated sentiments and the original ones. Finally, a comparative study of FEEL with the existing French sentiment lexicons have shown very competitive results for polarity and emotion classification on reference benchmarks. The FEEL lexicon has been made freely available to the community.

Our last contribution concerned the evaluation of different combinations of features and methods for French sentiment classification. Benchmarks released in previous sentiment classification challenges have been used in these evaluations. The considered text documents consisted in tweets and product reviews, while the considered classification tasks were polarity, subjectivity and emotion detection. We implemented the state of the art hand crafted features in sentiment analysis and included already learned Word Embeddings. Then, we proposed a feature engineering process that chooses at each step (pre-processings, syntactic features, word embeddings, etc.) the characteristics that improve the results with respect to chosen ones. This process has been applied by cross validation on the training set of each benchmark. The selected configuration have been used to learn appropriate classification models. When applied to the test sets, our models obtained comparable results with the best performing systems at each challenge. Furthermore, this study allowed us to find the features that are useful in sentiment classification of text documents from different nature and length. For example, lexicon based features are more useful in the classification of short texts, bigrams are useful in the classification of long text documents, etc. These findings may be considered to choose proper feature sets when learning French sentiment classification models. Finally, the learned systems and the implemented features have been made publicly available on a dedicated web-platform.

7.2 Future Research

During this thesis, several limitations and research directions have been identified and judged as deserving a deeper study. At the end of each chapter, short term perspectives have been presented dealing the two remaining contextual dimensions described in the introduction. For instance, we suggested to study the user expertise over time (when?), to compute a topic-dependent reputation value (what?), etc. In this section, we present some promising future research paths for the middle and long run.

7.2.1 User Mining by Combining Multiple Sources

Mining Social Media users has attracted research for different applications (Omidvar et al., 2014,Tagarelli and Interdonato, 2014,Li et al., 2015). Most of these studies do not transfer the knowledge extracted about each user between different Social Medias (e.g. Facebook, Twitter, Instagram, WhatsApp, Tumblr, Google+, Snapchat, StackExchange, Reddit, etc.). However, it will be interesting to combine multiple sources in order to have a global view of the same user over different Social Media. It is possible to exploit the knowledge extracted from one source to enhance the mining process in the remaining sources. In this process, user matching techniques are necessary (Liu et al., 2013,Korula and Lattanzi, 2014,Liu et al., 2014,Goga et al., 2015). These techniques are usually based on private and public profile attributes. User mining by combining multiple sources will have plenty of useful applications. For example, we will be able to compute the user expertise in diverse sources in a much more accurate way. Indeed, we may use the user expertise computed for a given website when the same user registers in another website to avoid the problem of cold-start. Similarly, we can take advantage from the topics that have been already discussed by the user in some websites to infer his expertise when the same topic is discussed in another website.

7.2.2 Multi-modal Social Media Mining

As mentioned in this thesis, combining Text Mining and Social Link Analysis allows a much better understanding of the social data. However, today's Social Media contents are much more diversified. It would be interesting to mine not just network structures and textual messages, but also pictures, speeches, videos, locations, etc. Research in Social Media has already started combining contents of different types. For instance, (Wang et al., 2015) conducted unsupervised sentiment analysis from large-scale Social Media images, considering both visual content and contextual information, such as comments and captions. (Yue-Hei Ng et al., 2015) applied Deep Neural Networks for videos classification. (Park and Yu, 2015) used Text Mining techniques in order to perform spatial clustering of location-based Social Media data. Mining these new types of contents can rely on the huge work done

in each domain separately: Text Mining, Network Analysis, Speech Recognition, Computer Vision, etc. Due to the rapid growth of these Social Media contents, it is expected that there will be much interest in Multi-modal Social Media Mining in the coming years. However, privacy issues may be even more important than now. Indeed, Social Media users may be more sensitive when it comes to their personal pictures or their real-time locations.

7.2.3 Enhancing Sentiment Analysis

Research in sentiment analysis is still facing many challenges and attracting tremendous applications (Mohammad, 2016). First, there is growing interest in detecting figurative language, especially irony and sarcasm (Rosenthal et al., 2015). Indeed, sarcasm and irony are very difficult to identify. The results of the sentiment classification models submitted to SemEval 2014 dropped by about 25 to 70 percent when applied to a separate test set involving sarcastic tweets (Rosenthal et al., 2014). Then, it has been reported that building specific models for each language induces better results than translating the textual documents to English and using the state of the art English models (Mohammad et al., 2015a). Indeed, cultural differences can lead to significantly different sentiment expressions. For this purpose, it is interesting to adapt the used features, methods and resources to each language. For instance, specific sentiment lexicons can be compiled for other languages following the approach proposed in chapter 5. Finally, recent research indicates that sentiment lexicons focusing on a specific domain leads to better sentiment classification results (Park et al., 2015). Therefore, it would be interesting to adapt the state of the art sentiment lexicons to the studied domain. Regarding the applications, they are phenomenally increasing in very different domains. Illustrating examples include: identifying the current public opinion towards the election candidates (Mohammad et al., 2015b), detecting personality traits (Grijalva et al., 2015), predicting health attributes (Eichstaedt et al., 2015), etc.

7.2.4 Ethical Considerations in Social Media Mining

Social Media Mining raises ethical issues on the way private and public personal data are being processed[1]. Some Social Medias such as Facebook allow their users to specify the persons who can access to their personal information and posted contents (friends, friends of friends, public, etc.). In these Social Medias, private data can not be accessed and therefore collected. However, even when the data is accessible, it is still problematic to decide whether it is private or not. The best way to ensure that the data can be used is to seek informed consent from Social Media users. However, most of the time, users do not read properly agreement forms and are not aware that the data they produce can be used for academic research or economic purposes. Indeed, acquiring informed consent is problematic in the case of datasets

[1] www.dotrural.ac.uk/socialmediaresearchethics.pdf

issued from Social Media (Hutton and Henderson, 2015). Moreover, anonymisation is a key consideration when sharing the used datasets or publishing qualitative results. This process may be complex (Narayanan and Shmatikov, 2009), especially when texts are associated with sound, image and videos. The goal behind studying these considerations is to reveal the conditions under which social datasets may be collected, processed and distributed. Indeed, distributing such common social datasets will ensure the reproducibility of the methods and avoid wasting time and resources in annotation steps.

Globally, my future research concerns the area of Social Data Science considering all the contextual dimensions presented in the introduction of this thesis (Who, How, What and When). It will explore the perspectives presented in this section and consider the issues of Big Data Analytics (Volume, Variety, Velocity, Veracity, etc.).

Bibliography

Abdaoui, A., Azé, J., Bringay, S., Grabar, N., and Poncelet, P. (2014). Predicting medical roles in online health fora. In *Proceedings of the International Conference on Statistical Language and Speech Processing*, pages 247–258. Springer.

Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2015a). Collaborative content-based method for estimating user reputation in online forums. In *Proceedings of the 16th International Conference on Web Information Systems Engineering*, pages 292–299. Springer.

Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2016). Feel: French expanded emotion lexicon. *Language Resources and Evaluation*, pages 1–23.

Abdaoui, A., Tapi Nzali, M. D., Azé, J., Bringay, S., Lavergne, C., Mollevi, C., and Poncelet, P. (2015b). Advanse : Analyse du sentiment, de l'opinion et de l'émotion sur des tweets français. In *Actes du 11e Défi Fouille de Texte*, pages 78–87.

Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM.

Agarwal, N., Liu, H., Tang, L., and Yu, P. S. (2008). Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining*, pages 207–218.

Ali, T., Schramm, D., Sokolova, M., and Inkpen, D. (2013). Can i hear you? sentiment analysis on medical forums. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 667–673.

Anjaria, M. and Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. In *Proceedings of the Sixth International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–8.

Asher, N., Benamara, F., and Mathieu, Y. Y. (2008). Distilling Opinion in Discourse: A Preliminary Study. In *Proceedings of the International Conference on Computational Linguistics*, pages 7–10.

Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., and Rinck, F. (2006). Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique.*, pages 407–414, Grenoble, France.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference*, volume 10, pages 2200–2204.

Balahur, A. (2013). Sentiment analysis in social media texts. In *4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 120–128.

Bekkerman, A. R. and McCallum, A. (2004). Extracting social networks and contact information from email and the web. In *Proceedings of the first Conference on Email and Anti-Spam*.

Brandes, U. and Erlebach, T. (2005). *Network analysis: methodological foundations*, volume 3418. Springer Science & Business Media.

Bringay, S., Kergosien, E., Pompidor, P., and Poncelet, P. (2014). Identifying the targets of the emotions expressed in health forums. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 85–97. Springer.

Che, D., Safran, M., and Peng, Z. (2013). From big data to big data mining: challenges, issues, and opportunities. In *Proceedings of the International Conference on Database Systems for Advanced Applications*, pages 1–15. Springer.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Cormack, G. V. (2007). Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cui, H., Mittal, V., and Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. In *Proceedings the twenty-first AAAI Conference on Artificial Intelligence*, volume 6, pages 1265–1270.

Dermouche, M., Velcin, J., Khouas, L., and Loudcher, S. (2014). A joint model for topic-sentiment evolution over time. In *2014 IEEE International Conference on Data Mining*, pages 773–778. IEEE.

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes.

Dray, G., Plantié, M., Harb, A., Poncelet, P., Roche, M., and Trousset, F. (2009). Opinion mining from blogs. *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, 1:205–213.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., et al. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2009). Improving the extraction of bilingual terminology from wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 5(4):31.

Farzindar, A. and Inkpen, D. (2015). *Natural Language Processing for Social Media*. Morgan & Claypool Publishers.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Forestier, M., Stavrianou, A., Velcin, J., and Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems: An international Journal*, 10(1):117–133.

Forestier, M., Velcin, J., and Zighed, D. (2011). Extracting social networks enriched by using text. In *International Symposium on Methodologies for Intelligent Systems*, pages 140–145. Springer.

Gala, N. and Brun, C. (2012). Propagation de polarités dans des familles de mots: impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions. In *Actes de Traitement Automatique des Langues Naturelles, Grenoble*, pages 495–502.

Goga, O., Loiseau, P., Sommer, R., Teixeira, R., and Gummadi, K. P. (2015). On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1799–1808. ACM.

Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web (TWEB)*, 3(4):12.

Grabar, N., Chauveau-Thoumelin, P., and Dumonet, L. (2016). Medical Discourse and Subjectivity. In Guillet, F., Pinaud, B., Venturini, G., and Zighed, D. A., editors, *Advances in Knowledge Discovery and Management*, number 615 in Studies in Computational Intelligence, pages 33–54. Springer International Publishing.

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., and Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological bulletin*, 141(2):261–310.

Grouin, C., Hurault-Plantet, M., Paroubek, P., and Berthelin, J.-B. (2009). Deft'07: une campagne d'évaluation en fouille d'opinion. *Fouille de données d'opinion*, 17:1–24.

Gruzd, A. A. and Haythornthwaite, C. (2008). Automated discovery and analysis of social networks from threaded discussions. In *Proceedings of the International Network of Social Network Analysis*.

Guille, A., Hacid, H., Favre, C., and Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28.

Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., and Ronen, I. (2013). Mining and interests from social media. In *Proceedings of the 22nd international conference on World Wide Web*, pages 515–526. ACM.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182.

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004). Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576–587.

Haddi, E., Liu, X., and Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, 17:26–32.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hamdan, H., Bellot, P., and Bechet, F. (2015). Sentiment lexicon-based features for sentiment analysis in short text. In *Proceeding of the 16th International Conference on Intelligent Text Processing and Compu- tational Linguistics.*

Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de texte (deft). In *Actes de la 11e Défi Fouille de Texte*, pages 1–11.

Hamon, T. and Nazarenko, A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents Web : retour d'expérience. *Traitement Automatique des Langues*, 49(2):127–154. 28 pages.

Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., and Poncelet, P. (2008). Web opinion mining: How to extract opinions from blogs? In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, pages 211–217. ACM.

Haykin, S. and Network, N. (2004). A comprehensive foundation. *Neural Networks*, 2(2004).

Heidemann, J., Klier, M., and Probst, F. (2010). Identifying key users in online social networks: A pagerank based approach. In *Proceedings of the International Conference on Information Systems.*

Himmel, W., Reincke, U., and Michelmann, H. W. (2009). Text mining and natural language processing approaches for automatic categorization of lay requests to web-based expert forums. *Journal of medical Internet research*, 11(3):e25.

Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM.

Hu, X., Tang, J., Gao, H., and Liu, H. (2014). Social spammer detection with sentiment information. In *Proceedings of the IEEE International Conference on Data Mining*, pages 180–189.

Hutton, L. and Henderson, T. (2015). " i didn't sign up for this!": Informed consent in social network research. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM)*, pages 178–187.

James, W. (1884). What is an emotion? *Mind*, (34):188–205.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Kardan, A., Omidvar, A., and Behzadi, M. (2012). Context based expert finding in online communities using social network analysis. *International J of Computer Science Research and Application*, 2(1):79–88.

Kardan, A., Omidvar, A., and Farahmandnia, F. (2011). Expert finding on social network with link analysis approach. In *Proceedings of the 19th Iranian Conference on Electrical Engineering*, pages 1–6. IEEE.

Kennedy, A. and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125.

Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8. Association for Computational Linguistics.

Kinsella, S., Passant, A., and Breslin, J. G. (2011). Topic classification in social media using metadata from hyperlinked objects. In *European Conference on Information Retrieval*, pages 201–206. Springer.

Kiritchenko, S., Mohammad, S. M., and Salameh, M. (2016). Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval), San Diego, California, June*.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50(1):723–762.

Klebanov, B. B., Madnani, N., and Burstein, J. (2013). Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics*, 1:99–110.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

Korula, N. and Lattanzi, S. (2014). An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment*, 7(5):377–388.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015a). *Games with a Purpose (GWAPS)*. John Wiley & Sons.

Lafourcade, M., Le Brun, N., and Joubert, A. (2015b). Vous aimez ?...ou pas ? likeit, un jeu pour construire une ressource lexicale de polarité. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, pages 330–336, Caen, France. Association pour le Traitement Automatique des Langues.

Lamirel, J.-C., Cuxac, P., Chivukula, A. S., and Hajlaoui, K. (2015). Optimizing text classification through efficient feature selection based on quality metric. *Journal of Intelligent Information Systems*, 45(3):379–396.

Lampe, C. and Resnick, P. (2004). Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550. ACM.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Larkey, L. S. (1999). A patent search and classification system. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 179–187. ACM.

Le, H.-S., Oparin, I., Allauzen, A., Gauvain, J.-L., and Yvon, F. (2011). Structured output layer neural network language model. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5524–5527. IEEE.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Leskovec, J., Lang, K. J., and Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World Wide Web*, pages 631–640. ACM.

Li, B., Li, R.-H., King, I., Lyu, M. R., and Yu, J. X. (2015). A topic-biased user reputation model in rating systems. *Knowledge and Information Systems*, 44(3):581–607.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Liu, J., Zhang, F., Song, X., Song, Y.-I., Lin, C.-Y., and Hon, H.-W. (2013). What's in a name?: an unsupervised approach to link users across communities. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 495–504. ACM.

Liu, Q., Li, J., Wang, Y., Xing, G., and Ren, Y. (2014). Account matching across heterogeneous networks. In *Proceedings of the 5th International Conference on Game Theory for Networks*, pages 1–5. IEEE.

Marsh, S. P. (1994). *Formalising trust as a computational concept*. PhD thesis, University of Stirling.

Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 301–311. Springer.

Matsuo, Y., Mori, J., Hamasaki, M., Nishimura, T., Takeda, H., Hasida, K., and Ishizuka, M. (2007). Polyphonet: an advanced social network extraction system from the web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(4):262–278.

Matsuo, Y., Tomobe, H., Hasida, K., and Ishizuka, M. (2004). Finding social network for trust calculation. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, volume 16, page 510.

McAuley, J. J. and Leskovec, J. (2013). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908. ACM.

Melzi, S., Abdaoui, A., Azé, J., Bringay, S., Poncelet, P., and Galtier, F. (2014). Patient's rationale: Patient knowledge retrieval from health forums. In *Proceedings of the 6th International Conference on eHealth, Telemedicine, and Social Medicine*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Mitchell, T. M. (1997). *Machine learning.*, volume 45. Burr Ridge, IL: McGraw Hill.

Mohammad, S. (2012). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Meiselman, H., editor, *Emotion Measurement*. Elsevier.

Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2015a). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 1:1–20.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2015b). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. Citeseer.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2016). Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18.

Narayanan, A. and Shmatikov, V. (2009). De-anonymizing social networks. In *Proceedings of the 30th IEEE symposium on security and privacy*, pages 173–187. IEEE.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Sentiful: A lexicon for sentiment analysis. *Affective Computing, IEEE Transactions on*, 2(1):22–36.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational linguistics*, 30(4):417–449.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 122–129.

Omidvar, A., Garakani, M., and Safarpour, H. R. (2014). Context based user ranking in forums for expert finding using wordnet dictionary and social network analysis. *Information Technology and Management*, 15(1):51–63.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: bringing order to the web.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference*, volume 10, pages 1320–1326.

Pal, A. R. and Saha, D. (2014). An approach to automatic text summarization using wordnet. In *IEEE International Advance Computing Conference (IACC)*, pages 1169–1173. IEEE.

Paltoglou, G. and Thelwall, M. (2012). Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):66.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL conference on Empirical methods in natural language processing*, pages 79–86. Association for Computational Linguistics.

Park, S., Lee, W., and Moon, I.-C. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, 56:38–44.

Park, W. J. and Yu, K. Y. (2015). Spatial clustering analysis based on text mining of location-based social media data. *Journal of Korean Society for Geospatial Information System*, 23(2):89–96.

Parrott, W. G. (2001). *Emotions in social psychology: Essential readings*. Psychology Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., and Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5(Suppl 1):3–16.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods*, pages 185–208. MIT Press.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 27–35. Citeseer.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 9, pages 1199–1204.

Qu, B., Cong, G., Li, C., Sun, A., and Chen, H. (2012). An evaluation of classification models for question topic categorization. *Journal of the American Society for Information Science and Technology*, 63(5):889–903.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Rafiei, M. and Kardan, A. A. (2015). A novel method for expert finding in online communities based on concept map and pagerank. *Human-centric computing and information sciences*, 5(1):1–18.

Rangel, F. and Rosso, P. (2016). On the impact of emotions on author profiling. *Information Processing & Management*, 52(1):73–92.

Rangel, F., Rosso, P., Potthast, M., Stein, B., and Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In *Proceedings of the Sixth Conference and Labs of the Evaluation Forum*, pages 50–58, Toulouse, France.

Rao, Y., Lei, J., Wenyin, L., Li, Q., and Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17(4):723–742.

Rastogi, S., Singhal, R., and Kumar, A. (2014). An improved sentiment classification using lexicon into svm. *International Journal of Computer Applications*, 95(1):37–42.

Riloff, E., Wiebe, J., and Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the national conference od the american association for artificial intelligence*, volume 20, pages 1106–1111. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Rokach, L. and Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487.

Rosa, K. D., Shah, R., Lin, B., Gershman, A., and Frederking, R. (2011). Topical clustering of tweets. In *Proceedings of the ACM SIGIR workshop Social Web Search and Mining Under Crisis*.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463.

Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80.

Rouvier, M., Favre, B., and Andiyakkal Rajendran, B. (2015). Talep @ deft'15 : Le plus cooool des systèmes d'analyse de sentiment. In *Actes de la 11e Défi Fouille de Texte*, pages 97–103. Association pour le Traitement Automatique des Langues.

Sadat, F., Yoshikawa, M., and Uemura, S. (2003). Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 141–144. Association for Computational Linguistics.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, volume 12, pages 44–49.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Son, L. H., Allauzen, A., and Yvon, F. (2012). Continuous space translation models with neural networks. In *Proceedings of the 12th conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 39–48.

Soukoreff, R. W. and MacKenzie, I. S. (2001). Measuring errors in text entry tasks: an application of the levenshtein string distance statistic. In *CHI'01 extended abstracts on Human factors in computing systems*, pages 319–320. ACM.

Stavrianou, A., Andritsos, P., and Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *ACM Sigmod Record*, 36(3):23–34.

Stavrianou, A., Velcin, J., and Chauchat, J.-H. (2009). Definition and measures of an opinion model for mining forums. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in*, pages 188–193.

Stone, P., Dunphy, D. C., Smith, M. S., and Ogilvie, D. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1):113–116.

Strapparava, C., Valitutti, A., et al. (2004). Wordnet affect: an affective extension of wordnet. In *Proceedings of the Language Resources and Evaluation Conference*, volume 4, pages 1083–1086.

Sztompka, P. (1999). *Trust: A sociological theory*. Cambridge University Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tagarelli, A. and Interdonato, R. (2014). Lurking in social networks: topology-based analysis and ranking methods. *Social Network Analysis and Mining*, 4(1):1–27.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)*, pages 1555–1565.

Tang, J., Chang, S., Aggarwal, C., and Liu, H. (2015). Negative link prediction in social media. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 87–96. ACM.

Tang, J., Gao, H., Liu, H., and Das Sarma, A. (2012). etrust: Understanding trust evolution in an online world. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 253–261. ACM.

Tang, J., Hu, X., and Liu, H. (2013). Social recommendation: a review. *Social Network Analysis and Mining*, 3(4):1113–1133.

Tapi Nzali, M. D., Bringay, S., Lavergne, C., Opitz, T., Azé, J., and Mollevi, C. (2015). Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. In *Actes des 26es journées froncophone d'Ingénierie des Connaissances*, pages 9–20, Rennes, France.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009). Mining Multi-label Data. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424.

Vincent, M. and Winterstein, G. (2013). Construction et exploitation d'un corpus français pour l'analyse de sentiment. In *Actes TALN-RÉCITAL*, pages 764–771.

Wanas, N., El-Saban, M., Ashour, H., and Ammar, W. (2008). Automatic scoring of online discussion posts. In *Proceedings of the 2nd ACM Workshop on information Credibility on the Web*, pages 19–26. ACM.

Wang, Y., Wang, S., Tang, J., Liu, H., and Li, B. (2015). Unsupervised sentiment analysis for social media images. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina*, pages 2378–2379.

Welser, H. T., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, 8(2):1–32.

Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., de Oliveira, J. P. M., and Wives, L. K. (2014). Examining multiple features for author profiling. *Journal of Information and Data Management*, 5(3):266.

Widén-Wulff, G., Ek, S., Ginman, M., Perttilä, R., Södergård, P., and Tötterman, A.-K. (2008). Information behaviour meets social capital: a conceptual model. *Journal of information science*, 34(3):346–355.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Yang, C., Lin, K. H.-Y., and Chen, H.-H. (2007). Emotion classification using web blog corpora. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pages 275–278. IEEE.

Yang, S.-H., Kolcz, A., Schlaikjer, A., and Gupta, P. (2014). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM.

Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3):6527–6535.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702.

Zafarani, R., Abbasi, M. A., and Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM.