Feature selection for opinion polarity detection by machine learning method

Fayçal R. Saidani[a], Idir. Rassoul[b]

[a] LARI laboratory—Mouloud Mammeri University, Tizi-Ouzou 15000, Algeria
saidani.faycal@gmail.com,

[b] LARI laboratory—Mouloud Mammeri University, Tizi-Ouzou 15000, Algeria
idir_rassoul@yahoo.fr

Abstract. Recent years have seen increasing interest in techniques of opinion mining and subjectivity analysis. In this article, we outline the results generated by our approach to detecting features for the classification polarity of opinions in French language using machine learning techniques. Indeed, in sentiment analysis, identifying features associated with an opinion can help to produce a finer-grained understanding of subjective previews. In this article, the proposed system consists of three phases: the pretreatments of the corpus, the extraction of the features and the classification. The second phase of our work represents the combination of the co-occurrence analysis for a better management of the intrinsic semantics of a word carrying opinion, and therefore a better extraction of features for classification.

Keywords: opinion mining, intrinsic semantics, polarity classification, Textometry, co-occurrence, Features extraction.

1 INTRODUCTION

Detection of opinions (also known as Sentiment Analysis) is the subject of a particular craze whether in academia or industry. Indeed, with the emergence of discussion groups, forums, blogs and compiling consumer reviews site, there is a very large mass of documents containing information expressing opinions, constituting a huge source of data for various applications survey (technological, marketing, competitive, societal). Much research at the crossroads of NLP and data mining, is addressing the problem of detecting opinions. The "bag of words" approach is one of the first models of textual representation, which is still today often used for sentiment analysis. The text is represented as a set of n-grams without consideration of their order of appearance and relationships in the text. Traditional approaches in machine learning (Naive Bayes or SVM) then use this representation to construct sentiments classification systems.

The accuracy of this kind of approach can be very high, especially when advanced features selection techniques are used in conjunction with additional lexicons extracted from texts previously identified as a carrier of opinion. However, we believe that model properties can identify more complex expressions of sentiments beyond simple recognition of opinionated construction, which should allow obtaining better classification systems. One problem with the bag of words approach is the loss of information during the construction of textual representation, seen as collections of differentiated terms. Yet, the relationship between the words in the text are often very important when determining whether the degree or the polarity of a sentiment.

In this article, we present the method used for the selection of textual features used by machine learning methods. For this, we proceeded to the preparation of data using tools of

textometric analysis to select those features. Many researchers have worked on the identification of features and sentiments. (Hu & Liu, 2004; Liu et al., 2005) proposed several techniques for mining feature-associated opinions expressed in reviews. In (Su et al., 2008) an unsupervised approach based on the mutual reinforcement principle is presented. Their approach clusters features and opinion words simultaneously by using both content and link information. To identify the feature-sentiment pair for real-life reviews, a new statistical Natural Language Processing approach that combines both syntactic and statistic knowledge was proposed in (Hai et al., 2010). These are used as a vector representation of the texts so that they can be used for supervised learning. To make this selection, we assume the following postulate: "to choose between positive or negative polarity of a text, we shall merely detect in it the indicators of opinions. Several methods can be used, for example the presence or absence of a set of determined words, the location of certain words (Hai et al., 2010), the identification of co-occurring subjectivity clues for the deduction of the overall character of the text, adjective (Saidani & Rassoul, 2013), adverb-adjective collocations and finally the syntactic dependencies (Nakagawa, Inui & Kurohashi, 2010). For our part, we believe that the presence of adjectives would be an indicator of this polarity.

The plan of the paper is as follows, after a brief description of the corpus used, we will present in the next section, the methodology for the creation and extraction of these characteristics. Finally, we present the results obtained from this selection process of features.

2 RELATED WORKS

The manner in which people express their opinions change depending on what they are talking about. The semantics of the words used is different from one area to another and an opinion classifier trained on a given field cannot be applied to extracts from another area without a minimum of adaptation. Thus, a word found in two different areas can easily change the intrinsic semantics (Navigli, 2012), and therefore lead to misclassification (Wilson et al., 2009). Most of the work about the intrinsic semantics aimed at improving an existing general lexicon, for example by giving polarities of words, depends on the weight field (Choi & Cardie, 2009). Also studies about classifiers based on corpus have mainly focused on the representation of data (Huang & Yates, 2012). Indeed, the adaptation error of a classifier depends on its performance in the source domain and the distribution difference between the source and target fields (Ben-David et al., 2007). With a good analysis of the context, we can establish links between words in the target domain that are absent from the source domain and the other (Pang & Lee, 2010; Blitzer et al., 2007). However, few studies focus on the recognition of the intrinsic semantics, a notable exception is (Yoshida et al., 2011), that uses a Bayesian formulation of the problem and focuses more specifically on the influence of the number of areas on the classification of opinions.

Concerning the extraction of features, different lexical features have been used in opinion mining. Various lexical features were compared in (Baccianella & Esuli, 2010). Unigrams and higher order n-grams are widely used. Word-meanings are used usually in their polar or emotional aspects, as for example in (Hannah et al., 2007). Adjectives, along with other features, such as parts of speech, syntax constructions, and the use of negation are the main classes of features used and compared in polarity classification; see (Esuli & Sebastiani, 2006) for examples. In (Pang & Lee, 2008) a hierarchy of lexical features is presented, the information gain of different features is discussed, and the hierarchy is employed for the selection of the best features for opinion analysis.

3 PRESENTATION OF THE CORPUS

The annotated corpus used for this experiment, is one of those, provided in the third edition of Text Mining Challenge 2007 (Grouin et al., 2007). The task requested from participants, was to classify texts according to their argument that could be rather positive, negative or neutral. The texts come from a variety of fields: film review, books, shows and comic strips, video game testing, proof reading of conference papers and parliamentary debates. We chose to work on a part of the corpus of the parliamentary debates. It includes 28 832 interventions of Deputies to the French National Assembly, extracted from debates on the Energy Law. Two values of opinion are available for this corpus: vote in favour of the law under consideration (positive class) and negative vote for the law under consideration (negative class). Only the interventions of more than 300 characters have been retained for the challenge, documents below this threshold were not considered usable after tests with human judges. However, with regard to the experience written in this article, we are restricted to only 2000 interventions equitably distributed according to the two categories of classes (positive, negative) and consequently we obtain two sub corpora (1000 interventions) for each respective category.

4 DESCRIPTION OF THE EXPERIMENT

4.1 Preprocessing Phase

These pretreatments consist in extracting linguistic units that will be used for the representation of texts in this corpus. This phase consisted of:

- Reduction of language by deleting the function words, not representative of an opinion. By function words, we mean the most frequent words of the French language in particular, *le la, un, une…* (List of 36 words).

- Reducing words to their lemmas to group words carrying the same meaning. Thus, for this experiment, a linguistic unit is considered as a lemmatized word.

This corpus analysis was performed using the Textometry software, TXM (Heiden et al., 2012) and Tree Tagger for the morph syntactic tagging.

4.2 Features selection phase and representation of the corpus

In this experiment, each category corresponds to a polarity of opinions (positive, negative). The final goal is to assign to each text one of these categories, reflecting the opinions expressed in this document. This is possible by searching chains of the characteristic words for each category, allowing to extract the comments.

The initial idea consists in taking into account changes in the intrinsic semantics of adjectives and this by separating the features that may most influence the learning phase. For this, we use a list of pivots words selected semi-automatically so that they do not change polarity. This selection is done in two stages: initially a preselection is performed so that the chosen words should not be too representative of a sub corpus and to be useful in the classification, thereafter an iterative process allows purifying this word list changing intrinsic semantics. First, we calculate the mutual information between the presence and absence of a word in an extract of the corpus and its membership in one of two sub corpora. This mutual information must be high, so that these words will be useful for the detection of polarity. Once the pivots words are selected, we perform the procedure to detect words changing semantics on the

pivots words themselves. This eliminates from the list the word most likely to change polarity. Then we start again until no word from the list is considered as semantically unstable.

These pivots words obtained and present in both sub corpora serve to compare the distribution of other adjectives in two respective sub-corpora. For each adjective, and for each corpus, we realize the profile of co-occurrence in relation to the list of pivots words. This research is based on specific algorithms and co-occurrence (Lafon, 1983) implemented in TXM. Thus, the analysis of co-occurrences focuses on contextual factors such as the direction and distance of lexical units to detect the overrepresentation of adjectives, and at the same time identify the lexical context attractions around this unit. This analysis uses statistical models among which are the Mutual Information and the hypergeometric model used by Lafon in the analysis of specific collocations implemented in TXM.

Subsequently a testing of Χ 2 allows us to determine if, for a given feature its co-occurrence profile in the source and target corpus is statistically different or not. We will keep ultimately a list of the first 30 adjectives as features for the vector representation.

4.3 Classification phase

The last phase of the experiment consists in giving to; each extract; one polarity (positive, negative) based on a supervised approach and having from the outset the documents constituting the corpus (Parliamentary debates). The two learning algorithms selected, were tested using TANAGRA (Ricco, 2005) a Data Mining tool for teaching and research that implements a set of methods resulting from the domain of exploratory statistics and machine learning.

5 RESULTS

The results of this experiment were evaluated by calculating the F-score of the corpus with β= 1, as shown in (Grouin et al., 2007) :

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (1)$$

The overall averages of the precision and recall of al l classes were evaluated according to the formula of the Macro-average:

$$Precision\ i = \frac{\text{Documents correctly assigned to the class } i}{\text{Number of documents assigned to class } i} \quad (2)$$

Order to compute the total accuracy; we calculate the average accuracies of each class:

$$Precision = \sum_{i=1}^{n} \frac{Precision_i}{n} \quad (3)$$

$$Recall_i = \frac{\text{Correctly documents assigned to the class } i}{\text{Number of documents belonging to class } i} \quad (4)$$

As for accuracy, we perform the average recall of each class to calculate the total recall:

$$Recall = \sum_{i=1}^{n} \frac{Recall_i}{n} \quad (5)$$

	Precision	Recall	F_{score}
Naïve bayes	0.695	0.637	0.664
SVM	0.723	0.691	0.706

Fig. 1. Execution results of Experiment.

We compared our results with interests of the state of art as shown in Fig. 1; our extraction procedure allows a selection of satisfactory characteristics in relation to the state of the art, which is encouraging. Also we can increase the precision without lowering the recall by refining the advantage of the filtering procedure of semantically unstable words. Indeed, we note that the method of screening pivot words plays a vital role in the performance, and in future work, we will study the optimal selection threshold and focus particular attention on the selection criteria.

6 CONCLUSIONS

In this paper, we propose a procedure for features selection for the detection of opinion. We have particularly taken into consideration changes in intrinsic semantics of words in the selection phase, this by purging our initial list of adjectives of all the elements that can influence the classification phase. The first results are encouraging and highlight the importance of the pre-selection of pivots words on the final results.

References

Baccianella, S., Esuli, A. (2010). *SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining*. In the conference on Language Resources and Evaluation, Valletta, Malta.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F. (2007). *Analysis of representations for domain Adaptation.* Advances in neural information processing systems, (vol. 19), 137.

Blitzer, J., Dredze, M., Pereira, F. (2007). *Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification.* In 45th annual meeting of the association for computational linguistics, Prague, Czech Republic.

Choi, Y., Cardie, C. (2009). *Adapting a polarity lexicon using integer linear programming for domain specific sentiment classification* . Conference on Empirical Methods on Natural Language Processing, Singapore.

Esuli, A., Sebastiani, F. (2006). *SentiWordNet : a publicly available lexical resource for opinion mining*. In Proceedings of Language Resources and Evaluation, Genova, Italy.

Grouin, C., Berthelin, J., Ayari, S. E., Heitz, T., Hurault-Plantet, M., Jardino, M., Khalis, Z., Lastes, M. (2007). *Présentation de DEFT 07*. In Text mining challenge 2007, Grenoble, France.

Hai, Z., Chang, K., Song, Q., Kim, J. (2010). *A statistical nlp approach for feature and sentiment identification from chinese reviews.* In proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, china, 105–112.

Hannah, D., Macdonald, C., Peng, J., He, B., Ounis, I. (2007). *Experiments in blog and enterprise tracks with terrier*. In Text Retrieval Conference, Maryland, USA.

Heiden, S., Magué, J-P., Pincemin, B. (2012). *TXM : Une plateforme logicielle open-source pour la textométrie*. In 10th international conference on statistical analysis of textual data, sapienza, Italy.

Hu, M., Liu, B. (2004). *Mining and summarizing customer reviews.* In proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, 342–351.

Huang, F., Yates, A. (2012). *Biased Representation Learning for Domain Adaptation.* , In proceeding of conference on empirical methods on natural language processing, Jeju Island, Korea, 1313-1323.

Lafon, P. (1983). *Analyse lexicométrique et recherche des cooccurrences*. Mots journal, (Vol. 3), 75-83.

Liu, B., Hu, M., Cheng, J. (2005). *Opinion observer: Analyzing and comparing opinions on the web.* In WWW 2005, Japan, 342–351.

Nakagawa, T., Inui, K., Kurohashi S. (2010). *Dependency tree-based sentiment classification using CRFs with hidden variables.* In proceedings of Human Language Technologies.

Navigli, R. (2012). *A quick tour of word sense disambiguation, induction and related approaches.* Theory and Practice of Computer Science, 115-129.

Pang, B., Lee, L. (2007). *Opinion Mining and Sentiment Analysis.* Foundations and Trends in Information Retrival, (vol. 2).

Pang, B., Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval, (Vol. 3), Issue 1–2.

Ricco, R. (2005). *TANAGRA : un logiciel gratuit pour l'enseignement et la recherche*. In Actes de EGC'2005, RNTI-E-3, vol. 2, 697-702.

Saidani, F.R.,Rassoul,I. (2013). *Cooccurrential analysis for a selection of discriminating features in opinion detection.* In international conference on information & intelligent systems, Sousse,Tunisia.

Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., Swen, B., Su, Z. (2008). *Hidden sentiment association in chinese web opinion mining.* In WWW in China-Mining the Chinese Web, Beijing, China, 959–968.

Wilson, T., Wiebe, J., Hoffmann, P. (2009). *Recognizing Contextual Polarity : An Exploration of Features for Phrase-Level Sentiment Analysis.* Computational Linguistics.

Yoshida, Y., Hirao, T., Iwata, T., Nagata, M., Matsumoto, Y. (2011). *Transfer Learning for Multiple-Domain Sentiment Analysis.* In AAAI 2011.