Opinion Mining Issues and Agreement Identification in Forum Texts

Anna Stavrianou — Jean-Hugues Chauchat

Université de Lyon
Laboratoire ERIC - Université Lumière Lyon 2
5 avenue Pierre Mendès-France
69676 Bron Cedex

Anna.Stavrianou@univ-lyon2.fr
Jean-Hugues.Chauchat@univ-lyon2.fr

ABSTRACT. *Opinion Mining refers to the identification of opinions and arguments in a text. Recently, it has received great attention due to the abundance of opinion data that reside in online discussions, reviews and conversational texts. In this paper, we study the challenges of Opinion Mining together with the published techniques and methodologies and we evaluate a method for detecting agreement or disagreement in a text. The method is still at its early stage and its originality lies in the fact that it attempts to find out agreement or disagreement statements as opposed to most current approaches that deal with positive, negative or neutral statements.*

RÉSUMÉ. *La fouille des données d'opinion (Opinion Mining) désigne les méthodes d'identification des opinions et argumentations au sein d'un ensemble de textes. Depuis peu, les recherches sur ce sujet se développent face au volume des textes d'opinions produits dans les discussions online, les commentaires sur des produits ou services, et les "chats". Dans cet article, nous présentons un état des méthodes et techniques publiées, puis nous évaluons une méthode pour détecter l'accord ou le désaccord dans un texte. La méthode en est à ses débuts et reste à perfectionner ; son originalité est de tenter d'identifier dans les textes des points d'accord ou de désaccord, contrairement à de nombreuses études actuelles qui recherchent des phrases à connotation positive, négative ou neutre.*

KEY WORDS: *Opinion Mining, Semantic Orientation, agreement, disagreement, forum, blog.*

MOTS-CLÉS: *Fouille des opinions, Orientation Sémantique, accord, désaccord, forum, blog.*

1. Introduction

Opinion Mining (OM) is an area of Text Mining that has recently received a lot of attention due to the amount of opinion information that resides in web documents. It concerns the identification of opinions in a text and their classification as positive, negative or neutral. Opinion identification is more difficult than the topic-based one and it cannot be based on just observing the presence of single words. More sophisticated methods need to be employed in order to differentiate between the subjective and objective opinion of a reviewer or between the objective description of a movie and references to other people's comments (Stavrianou *et al.,* 2007).

The identification of the opinion orientation can be manual, corpus-based and dictionary-based (Liu 2007). The manual one requires a lot of human effort and it is costly. The corpus-based one considers syntactic and statistical properties such as word co-occurrence, and it faces the problem of being domain dependent. The dictionary-based approach uses hierarchies and ontologies such as WordNet (http://wordnet.princeton.edu/) in order to identify the sentiment orientation but it faces the problem of the lack of context information in these hierarchies (Liu 2007).

Here, we present research and challenges in the field of OM. Then, we attempt to apply the logic of the existing techniques to the task of agreement analysis in forum texts. To our knowledge, the task of agreement identification has not been dealt with until now. Most of existing research regards the identification of opinion orientation. We apply the logic of connotation identification to the task of agreement identification by using a similarity measure between word features and an agreement/disagreement seed list. Our research shows that different techniques are needed in order to capture agreement in a text.

In the paper Section 2 discusses the various research challenges in OM while Section 3 presents a part of the nowadays state-of-art in the field. In Section 4 we evaluate an initial approach based on current techniques that attempts to identify whether a text shows agreement or disagreement. Section 5 concludes.

2. Challenges of Opinion Mining

The field of OM is recent and as a result there are still a lot of challenges to be met. According to (Liu 2007), current techniques are still primitive for both opinion and comparison identification and extraction.

The mining of forums and online discussions is a challenge on its own. The use of colloquial language, the abbreviations, the orthographic mistakes, the fact that the comments are entered by various people who differ in the way they write or in the knowledge of the language they use, are examples that show the difference of a forum text compared to a text that appears in an online newspaper.

An interesting issue is to be able to monitor how opinion changes over time. This will allow observing whether a product improves as the time passes, whether people become more satisfied with certain services, or even whether people are finally convinced after a long discussion in a forum. A research work which identifies how the mood of people changes over time – but not necessarily their opinion - is that of (Balog *et al.,* 2006) who observe blogs where the mood is explicitly specified either by selecting from a predefined list of moods or by entering it as free text.

Another challenge for OM is to study the strength of the identified opinions (Hu *et al.,* 2004). If a strong opinion can change during a discussion in a forum, it means that the arguments used have been strong as well. An attempt to identify opinion strength is done in the SentiWordNet application (Esuli *et al.,* 2005b).

The existence of sarcastic and ironic statements in a text cannot be identified currently and it may lead to erroneous orientation assignment and misleading opinion mining. (Turney *et al.,* 2003) highlight the importance of context, since a positive word may have a negative meaning in a metaphorical or ironic context.

Another challenge that we have attempted to initialize in this paper is the identification of agreement as well as the detection of which reactions are provoked by what statements. Existing research regards the discovery of which product feature an opinion refers to, but not on which idea an opinion word or phrase is expressed.

3. Existing Research

(Hatzivassiloglou *et al.,* 1997) are the first to deal with opinion classification. They focus on adjectives and they study phrases where adjectives are connected with conjunction words such as “and” or “but”. They construct a log-linear regression model so as to clarify whether two adjectives have the same orientation. Then they perform clustering to separate the adjectives into two classes, and they assume the cluster with the highest frequency to be the positive orientation cluster.

(Turney *et al.,* 2003) extract 2-word phrases where one word is an adjective or adverb. They use *pointwise mutual information* (PMI) and *latent semantic analysis* (LSA) to measure the relation between a word and a set of positive or negative words. The sum of the measurement between a word and the words from the positive seed set is subtracted from the sum of the association between the same word and the words in the negative set. The review is classified according to the average orientation of its phrases. The LSA-based measure gives better results.

(Kamps *et al.,* 2004) focus on adjectives. They follow the same logic as the work of (Turney 2002) with the difference that they use WordNet to define the semantic distance between the adjectives of a text and a set of already tagged words. The distance is defined as the path length between two graph nodes. They calculate the distance of a word from both “good” and its antonym “bad” and they propose three measures; the evaluative measure (“good”/“bad”), the potency measure (“strong”/“weak”) and the activity measure (“active”/“passive”).

(Pang *et al.,* 2002) classify sentiments of movie reviews. Their experiments show that machine learning algorithms, that give good results in thematic categorization, do not perform as well for sentiment classification. They show that the presence or absence of a word seems to be more indicative of the content rather than the frequency of a word. This has also been pointed out in (Wei *et al.,* 2008).

(Wiebe 2000) distinguishes between objective and subjective sentences. A seed set is constructed by tagging the subjective adjectives of a corpus and assigning a 1-3 score to each of them. For each adjective of score 3, 20 near-synonyms are found by using the similarity measure (Lin 1998) or WordNet. Semantic orientation features are added. The results show that a sentence is subjective by 55.8% if it contains at least one adjective. Also, the sentences that contain an adjective from the expanded seed set and the list of positive adjectives are subjective by 71%. It is claimed that ontologies and dictionaries are not sufficient to help distinguishing between facts and opinions because they are not tagged with subjectivity.

(Hu *et al.,* 2004) deal with product reviews from web sites in order to produce a summary with positive and negative statements made for product features. The model they use is presented in (Liu 2007). They identify features discussed in reviews (e.g. camera size, image etc.) by selecting the frequent words, assuming that people often use the same words to describe features. Then they identify opinion sentences and their orientation. An opinion sentence is defined as a sentence that contains both a feature and at least one adjective. They use a seed list of 30 basic adjectives. For each adjective in the reviews, they check whether it is in the seed list or it is an antonym or synonym of a word in the seed list. Every time the orientation of an adjective is found, the seed list is expanded with this adjective. The infrequent features are identified by looking for the nearest noun phrases to an opinion word. Finally each sentence is given the orientation of the majority of its part-orientations.

(Ding *et al.,* 2007) improve the previously mentioned (Hu *et al.,* 2004) system by assigning an orientation score to each opinion word found in a sentence. The score takes into account the semantic orientation of the opinion word that is located near the feature-word and the distance between the feature and the opinion word. In this way a low score is given to the opinion words that are far from the feature.

(Esuli *et al.,* 2005a) present a method that outperforms the results of known methods. This method is based on the assumption that terms with similar orientation tend to have similar glosses. The terms of the text are presented as vectors of glosses and they are weighted by tf-idf. They start with the seed set of (Kamps *et al.,* 2004) and (Turney *et al.,* 2003) that is enriched with the use of a thesaurus. Their research has resulted in SentiWordNet (Esuli *et al.,* 2005b), a lexical resource where each WordNet synset is associated to a score describing how positive, negative or objective the particular synset is. The score is the proportion of a committee of classifiers used in order to label every synset. This resource is still under evaluation.

Lately, (Ku *et al.,* 2007) have written an extensive review on OM for both Chinese and English news articles and blogs. They emphasize the fact that the extracted opinions are useful only if they are relevant to the topic and this is why they perform topic detection as an initial step of their approach. Subsequent to this step they extract sentences that contain opinion words. The opinion tendency of a word is calculated by a formula that takes into account language characteristics.

In Table 1, the approaches mentioned are presented. The majority focus on adjectives and adverbs. They use a seed list and they attempt to find out the relation between the words that appear in a text and the words of the seed list. The difference lies in the similarity measure used to calculate the association between words. Some use WordNet, others use statistical measures. Some approaches give importance to the percentage of how positive or negative a word is.

Approach	**Seed List**	**Score**	**Details**
(Ding *et al.*, 2007)	Same as (Hu *et al.*, 2004)	Distance of opinion words from features.	Objective: orientation of product reviews.
(Esuli *et al.*, 2005a, Esuli *et al.*, 2005b)	(Kamps *et al.*, 2004) and (Turney *et al.*, 2003) lists expanded by WordNet.	Scores decided by classifiers.	"SentiWordNet": scores to each WordNet synset.
(Hatzivassiloglou *et al.*, 1997)			Use categorization to detect orientation of conjoined adjectives.
(Hu *et al.*, 2004)	Initial seed list of 30 adjectives. Expanded by use of WordNet.		Identify frequent and infrequent features.
(Kamps *et al.*, 2004)	{good, bad}	Path length in WordNet between adjectives and the seed list.	Introduction of evaluative measures.
(Ku *et al.*, 2007)	Use thesaurus and dictionaries for seed list.	Formula for defining opinion tendency of a word.	Perform topic detection before opinion extraction.
(Pang *et al.*, 2002)			Absence/presence of words rather than frequency.
(Turney *et al.*, 2003)	{good, nice, excellent, positive, fortunate, correct, superior}, {bad, nasty, poor, negative, unfortunate, wrong, inferior}.	Statistical association (PMI, LSA) of a word with the words of the seed list.	Extraction of adverb, adjective phrases. LSA results were better than the PMI ones.

(Wiebe 2000)	List of adjectives of score 3 expanded by WordNet / similarity measure.		Distinction between subjective and objective phrases.

Table 1. *Current Approaches*

4. Identification of agreement or disagreement

We focus on the OM of texts that appear in forums, where people discuss about a particular subject. Our interests lie in discovering whether someone is for or against a statement, hence we orientate OM towards agreement and disagreement analysis.

To our knowledge, most existing work is about finding out the semantic orientation of a text, and not whether a text includes agreement or disagreement. Our objective is to connect opinion words with arguments and not with features, as in (Ding *et al.,* 2007 and Hu *et al.,* 2004). We are interested in discovering which positions are stated on what statement, in other words which reaction there is against what event. Moreover we envisage mining forums and debates rather than reviews.

Determining the orientation of a text does not reveal agreement or disagreement since someone may agree with a negative statement or disagree with a positive one. For example: “I think that Mr Smith is right (*agreement statement*) when he says that the right to keep and bear arms is dangerous (*negative connotation*)”. On this basis, our work is two-fold: a) point out agreement or disagreement and b) discover the statements on which it is expressed. Here, we deal only with the first part.

We evaluate an approach to mine forums in English. We consider the forum as a set of comments. The approach is at its early stage and it tests whether existing techniques work for agreement analysis. The method is divided into 3 main steps:

a) *Representation of the text as a set of words.* Initially we tokenize and tag each text by applying the TreeTagger software (Schmid 1994). We consider that agreement is expressed mainly with verbs (e.g. “agree”) and nouns (e.g. “agreement”), rather than adjectives. As a result, the words used in the representation are verbs or nouns, after stopword removal. Negation is also taken into account by replacing words preceded by negation with their WordNet antonyms - if they exist -. We realize that negation is often implicit but we just intend to capture statements as “I do not agree” and replace them with “I disagree”.

b) *Construction of a seed list that contains words that show agreement or disagreement.* We use as a seed list the words that appear in the General Inquirer (http://www.wjh.harvard.edu/~inquirer) categories that show agreement and disagreement. We restrict this list to only those words that appear in WordNet and we have expanded the list by adding synonyms and antonyms of each of the initial words, found in WordNet again.

c) *Calculation of the distance between the set of text words and the seed list.* In order to find out the distance between the two lists of words, we use the WordNet-Similarity package (http://www.d.umn.edu/~tpederse/similarity.html) that applies various similarity measures. We use the distributional similarity described in (Lin 1998), similarly to (Wiebe 2000). Each word that is present in the set of text words is compared against the words of the seed list. Two maximum similarity values are estimated for each word; one for agreement and one for disagreement. Each comment in the forum is assigned a score of agreement or disagreement according to how many words are more similar to agreement than disagreement and vice versa.

We have applied the method to a forum from www.englishforums.com, discussing the problems of having an accent when speaking English. We have extracted 30 comments in total from the forum that contain 77 words on average.

The method identifies agreement easier than disagreement, but the accuracy, when compared against human judgement, shows that the current techniques need to be enriched when they are applied to the task of agreement identification. The problem is that agreement is often implicit and not explicitly stated by specific words that could be captured by a seed list. Additionally, the language used in forums is satisfying comprehensible and not literature needs and, as a result, certain words and expressions are not always included in dictionaries such as WordNet.

In addition, not all forum comments express agreement or disagreement. In many of them questions, new ideas and stories are presented without necessarily having to refer to something previously said. This means that techniques that involve checking the similarity between a text and a seed list of words is not sufficient. The problem may be addressed by varying the similarity measures used. The knowledge we could acquire by identifying positivity or negativity in a text could also be used in having an idea about the orientation of the text before attempting to see its agreement state.

5. Conclusion

Opinion Mining is a rapidly growing field that deals with the identification of opinions and their orientation in a text. It is a field that poses a lot of challenges.

In this paper, we describe existing research and we apply existing techniques to the identification of agreement in a text. This is done by checking the similarity of text words against an agreement/disagreement seed list. The results show that identifying agreement features in a text is not a simple task and needs further and more advanced techniques. We plan to address this issue in the near future.

6. Acknowledgements

We thank all the anonymous reviewers for their valuable comments. This work has been funded by the Région Rhône-Alpes.

7. References

Balog K., Mishne G., De Rijke M., “Why are they excited? Identifying and explaining spikes in blog mood levels”, *EACL,* 2006.

Ding X., Liu B., “The utility of linguistic rules in Opinion Mining”, *SIGIR-07,* 2007.

Esuli A., Sebastani F., “Determining the semantic orientation of terms through gloss classification”, *CIKM-05,* 2005a, p. 617-624.

Esuli A., Sebastani F., “SENTIWORDNET: A publicly available lexical resource for Opinion Mining”, *LREC,* 2005b.

Hatzivassiloglou V., Mckeown, K.R, “Predicting the semantic orientation of adjectives”, *35th ACL and the 8th Conference of the European chapter of the ACL*, 1997, p. 174-181.

Hu M., Liu B., “Mining and summarizing customer reviews”, *KDD-2004,* 2004.

Kamps J., Marx M., Mokken R.J., De Rijke M., “Using WordNet to measure semantic orientations of adjectives”, *4th LREC*, 2004, p. 1115-1118.

Ku Lun-Wei, Chen Hsin-His, “Mining opinions from the web: beyond relevance retrieval”, *Journal of the American Society for Information Science and Technology*, 58(12): 2007.

Lin D., “Automatic retrieval and clustering of similar words”, *COLING-ACL,* 1998.

Liu B., “Tutorial on sentiment analysis” based on Chapter 11 of the book “*Web Data Mining – Exploring Hyperlinks, Contents and Usage Data*”. (http://www.cs.uic.edu/~liub/), 2007.

Pang B., Lee L., Vaithyanathan S., “Thumbs up? Sentiment classification using machine learning techniques”, *EMNLP,* 2002, p. 79-86.

Schmid H., “Probabilistic part-of-speech tagging using decision trees”, *International Conference of new methods in language processing*, 1994.

Stavrianou A., Andritsos P., Nicoloyannis N., “Overview and semantic issues of Text Mining”, *SIGMOD Record*, 36(3), 2007, p. 23-34.

Turney P.D., “Thumbs up or down? Semantic orientation applied to unsupervised classification of reviews”, *2002 ACL*, 2002, p. 417-424.

Turney P.D., Littman M.L., “Measuring praise and criticism: inference of semantic orientation from association”, *ACM TOIS 21(4)*, 2003, p. 315-346.

Wei Z., Chauchat J-H., Miao D., “Comparing different text representation and feature selection methods on Chinese text classification using character n-grams”, *JADT-2008*, 2008.

Wiebe J.M., “Learning subjective adjectives from corpora”, *AAAI-2000,* 2000.