Challenges for Sentence Level Opinion Detection in Blogs

Malik Muhammad Saad Missen
Université de Toulouse
IRIT UMR 5505 CNRS
118 Route de Narbonne
Toulouse France

Mohand Boughanem
Université de Toulouse
IRIT UMR 5505 CNRS
118 Route de Narbonne
Toulouse France

Guillaume Cabanac
Université de Toulouse
IRIT UMR 5505 CNRS
118 Route de Narbonne
Toulouse France

{missen, bougha, cabanac}@irit.fr

Abstract—**The Opinion Detection from blogs has always been a challenge for researchers. However with the introduction of Blog track in TREC 2006, a considerable improvement has been seen in this field at document level. But now it is the time when researchers are thinking to shift their orientation from opinion finding at document level to opinion finding at sentence or passage level. In this paper, we investigate the challenges the researchers might face with sentence-level opinion detection and have tried to demonstrate them with few examples. Our work also includes annotation of a small set of opinionated sentences by two annotators. These Annotators annotate the sentences by labels *Positive* or *Negative*. The results of annotation prove that task of opinion detection on sentence-level is more challenging task than opinion detection on document level. In addition, we also discuss the importance of sentence-level opinion detection. Our work can give a new direction to researchers to think and work on.**

Keywords- opinion; sentiment; information retrieval; Blogs; Polarity

I. INTRODUCTION

Opinion Detection is of one of the most exciting and challenging application of text analysis today. It is the ability of recognizing and classifying opinionated text within the documents [1]. This ability is desirable for various tasks, including filtering advertisements, separating the arguments in online debate or discussions, ranking web documents cited as authorities on contentious topics, etc.

In Opinion Detection, one has to check whether a given text has a factual nature (i.e. describes a given situation/event without giving any opinion about it) or expresses an opinion on its subject matter. This task can be performed on different levels of granularity, i.e. on word level, sentence level or on document level. As a conclusion of this task a given word, sentence or document can be declared as of opinionated nature (or subjective) or of factual nature (objective). Text with opinionated nature can further be analyzed for having negative or positive polarity of opinion and this subtask is called *Opinion Polarity Detection*. The task of Opinion Detection becomes more difficult and challenging when it is to be performed for Blogs. A blog (a contraction of the term Web log) is a Web site, usually maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Entries are commonly displayed in reverse-chronological order. "Blog" can also be used as a verb, meaning to maintain or add content to a blog [2]**.** Generally in blogs, bloggers (the authors of blog) express their opinions and thoughts about something being discussed in a blog. As we have discussed above, the task of Opinion Detection is an articulation of the information need that aims to uncover the public sentiment towards a given target entity such as a product, an organization or a location [3]**.** Therefore, blogs are the best source of opinions on the Web. There are several commercial blog search engines[1] that allow users to find out the opinions and thoughts of other people, who happily share their thoughts in blogs. These thoughts range from anger at some products, politicians or organizations, to good reviews of products or appraisal of cultural events.

Considering the importance of blogs regarding the task of Opinion Detection, TREC introduced a Blog track in 2006 known as TREC Blog Track with the release of blog data collection. The data collection is 148GB in size. This blog track is very important for evaluating a system's performance on the test blog data collection. Different groups with their systems have participated in TREC Blog 2006 and TREC Blog 2007. After analyzing the approaches used by TREC Blog participants, it can be noticed that there are two major approaches used for *Opinion Detection* i.e. Lexicon based approaches [4, 5, 6, 7] and Machine Learning based approaches [8, 9].

[1] http://www.blogsearchengine.com/
http://blogsearch.google.com/
http://www.blogdigger.com/

- In Lexical based method, entries are tagged with their positive/negative prior polarity, i.e. no context is taken into account. A final opinion score is calculated on behalf of polarities of the words (within context or without context). For example, *beautiful* has a positive prior polarity, and *horrid* has a negative prior polarity. However, the contextual polarity of the phrase in which a word appears may be different from the word's prior polarity. For example, if a positive word like *strong* is preceded by a negation *not (i.e. not strong)*, it inverses its prior polarity. There are many other features that can be used to determine the contextual polarity of a word [10]. There are various lexical resources [11, 12] available for this task.
- In Machine learning methods, usually a classifier is trained using a set of annotated texts containing sentiment, typically employing features such as n-grams of words, part-of-speech tags, and logical forms. The details about the use of these and some other approaches can be consulted in overview papers of TREC 2006 [13] and TREC 2007 [14]. This paper focuses on the need of work on sentence-level opinion detection and then discusses the challenges and applications of sentence-level opinion detection in blogs.

This paper is organized as follows: Section II discusses the importance of sentence-level opinion detection and in the following section III we discuss the challenges of sentence-level opinion detection in blogs. In section IV, we outline some avenues for research which also concludes the paper.

II. WHY SENTENCE-LEVEL OPINION DETECTION?

The main task of the TREC Blog is to detect the opinionated documents having opinions about a given target. So the basic idea is that if a document contains some words or sentences about the target (or says Query Words) and if these words or sentences are opinionated, the document is delivered to the user considering it a relevant opinionated document. Here, the question arises that, "Does the document only talks about the target present in the query?" The answer is not always "YES". The document might contain opinion about the target, some characteristic of the target not asked in query or any other topic or target. Like a review on a product review site may be positive as a whole but might contain some negative opinion about one of its features. For instance, a review about a digital camera may be positive as a whole but can contain few negative words or sentences about its optical view finder. So if someone wants to have feature-based positive or negative views about a product then documents have to be treated on sentence or passage level. In short, we can say that sentence-based opinion detection is very important for feature-level characterisation of products.

Further, when dealing on sentence level things become more contextual than on document level. The sense of words becomes more contextual and the sentences become more meaningful than on document level. For example, the series of few consecutive sentences can be helpful to resolve the contextual ambiguity of a word which is one of the main problems of document level opinion detection.

III. CHALLENGES

In the previous section, we have discussed the importance of work regarding sentence-level opinion detection. The realisation of this work is to face some challenges that are to be discussed in this section. We will discuss them one by one as follows:

A. Sentence Extraction

The foremost challenge for sentence-level opinion detection in blogs is to well recognise the sentence boundaries and split up the text of a blog in proper complete sentences. However sentence boundary detection is a challenging task. At first glance, it may appear just a usage of few short listed sentence-final punctuation marks, such as ".", "?", and "!" but it is to be noted that graphemes[2] often serve more than one purpose in writing systems. The punctuation marks are not used exclusively to mark sentence breaks. For example, embedded quotations may contain any of the sentence ending punctuation marks and a period is also used to mark abbreviations, initials, ordinal numbers, decimal points, dates, email, web site addresses and ellipses. Moreover, punctuation marks may be used to mark an abbreviation and a sentence boundary at the same time, or they may be used multiple times for emphasis to mark a single sentence boundary. Sentence boundary detection thus could be considered as an instance of ambiguity resolution [15, 16, 17]. All above problems become more severe when we have to deal with a blog data collection. To add with above problems, the challenges for sentence boundary detection while working with blogs go as below:

I) Blogs contain less information and more advertisement data because blogs proved to be one of the best sources of advertisement for commercial products [18]. According to [18], bloggers are vocal and influential so an influential blog is the best platform for a new brand to get popular in public that is supposed to be audience of a blog. This fact makes the task of sentence extraction more difficult because unnecessary data of advertisements can be mixed up with necessary information.

In Fig.1, blogposts within a blog are surrounded by advertisements. The advertisements are marked with bold rectangles.

[2] a grapheme is the fundamental unit in written language.

Figure 1. Example of a blog with multiple posts and ads

II) The use of informal language in blogs with no or rare use of punctuations makes the task of sentence extraction more difficult. Due to lack of proper use of punctuations, sentence splitters end up with a sentence which is either a concatenation of two (or more) independent sentences or one half of a complete sentence. Here is an example of a sentence extracted by using a sentence splitter on a blogpost:

"Yes his name is Tom please leave your comment email will not be published"

In this example, three sentences have been merged by sentence extractor to form one complete sentence. There is nothing wrong with sentence extractor but it happened because of lack of necessary punctuations in between sentences. Similarly another example of indifference while writing in blogs is given below:

"hahaha ... what the hell you are talking about. r u fooooooool?"

Text like above in the example is very common in blogs. As evident from this text, one cannot see capitalization and cannot distinguish between periods ending the sentence and the ones being used to mark abbreviations.

B. The Language of Sentences

Bloggers are not professional writers. They belong to all ages, cultures, regions and religions. They vary in their capabilities to write in English (or any other language in which they are writing). Therefore they may not follow the language grammar rules while writing a blog or blog comments. They try to use words present in their vocabulary in a way that is enough to convey their message. Even people proficient in English do not write in a newspaper style but they are used to follow Netspeak [19] (aka *Internet Slang*). *NetSpeak* is a process of shortening words and replacing letters with different letters and/or symbols to supposedly make the typing process shorter. It lacks proper punctuation and capitalization. Sometimes it is really hard to understand the real message behind such characters/or symbols. For example, let's have a look on a comment found on a video blog site:

"yeah man cooolest video ever gud work :)"

The use of emoticons like ☺ ☹, short words like *gud* inspite of *good,* misspelled words like *cooolest* in spite of *coolest* makes this sentence a difficult candidate for text processing. In addition, bad use of grammar and no use of punctuations make the task of text processing more challenging [19, p. 127].

In such a scenario, the formal methods of Natural Language Processing like parsing tree etc may not be as effective as in other cases. It is a big challenge especially in the task of opinion polarity detection where the effects of positive (like good, excellent, ideal etc) or negative (like not, never, no, bad, problematic, etc) words prolong a bit further in the sentence.

C. Sentence Polarity Detection

Another challenge that researchers have to face is the task of polarity detection for a sentence i.e. whether a sentence is positive or negative in its expressiveness. Things are relatively easy on document level because we decide about its polarity from top level. In this subsection, we will discuss some points that become more difficult when being dealt on sentence level polarity detection.

We start with the first challenge of sentence level polarity detection. On sentence level sometimes for some sentences, it becomes very difficult even for a human being to decide about its polarity after analysing its constituent parts**.** One example of such sentence is given below:

"I know he is not a good boy but he is not that bad too"

It is obvious from this sentence that one can classify it as positive or negative sentence. To deeply observe this phenomenon, we planned to observe Inter-Annotator Agreement (IAA) by annotating a collection of such sentences.

I) Annotation Study: We prepared a small set of 100 such sentences and asked two persons to annotate those sentences as positive or negative. None of the annotators were author of these sentences. One half of these 100 sentences were taken from TREC Blog collection and rest were chosen from different blogs sites[3] just to generalise our results for various topics. Table 1 shows the resulting Inter-Annotator agreement between the two annotators. In the instructions to annotators, we asked them to annotate

[3] http://www.huffingtonpost.com
http://www.youtube.com
http://www.wize.com

each sentence with label Positive or Negative. They were asked to label the words with their best choice.

TABLE 1. INTER-ANNOTATOR AGREEMENT SCORE

Annotator 2	Annotator 1		
	Pos	Neg	Total
Pos	21	22	43
Neg	17	40	57
Total	38	62	100

We have used the Cohen's Kappa measure [21] as Inter-Annotator agreement measure. Cohen's Kappa coefficient is a statistical measure of inter-annotators agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two annotators who each classify N items into C mutually exclusive categories [20]. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (1)$$

Where *Pr (a)* is the relative observed agreement among annotators, and *Pr (e)* is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement among the annotators (other than what would be expected by chance) then $\kappa \leq 0$.

We hypothesized that such sentences are very difficult to annotate with Positive or Negative and it is very hard for both annotators to get a high levels of agreement. The labels that both annotators assigned to sentences support our hypothesis. According to equation 1, the value of κ results in 0.20 which interprets to slight Inter-Annotator Agreement between the annotators hence proves our point. The interpretation of different values of κ [21] falling in the interval [-1, 1] is shown in table 2.

TABLE 2. KAPPA VALUE INTERPRETATIONS

Kappa Value (*k*)	Interpretation
Below 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

The value of κ calculated during annotation process is surprising. Such a low score not only proves our point regarding the difficulty of polarity task on sentence level but also gives an idea about the complexity of evaluation process of such systems. Therefore, we think that building an evaluation framework for sentence level opinion detection is another challenging task and need to be worked on.

The second challenge for sentence level polarity detection is polarity metric. In document level polarity detection, one of the most successful metric for polarity of a document is number of positive and negative words present in that document [22]. This metric cannot be effective on sentence level. Sentences contain relatively less number of words than documents. Having a few extra positive or negative words in a document might not affect the polarity of that document because of its scope but in a sentence, an extra positive or negative word can easily reverse the polarity of that sentence . In other words, sentences are more vulnerable to change in their polarity regarding the number of positive or negative words they contain and this situation goes worst when sentences are malformed or when unnecessary data gets mixed up with value data. In such a scenario, more effective and robust metrics are needed to be proposed for sentence level polarity detection.

IV. CONCLUSION

In today's modern world when online communities are becoming more social and informative, researchers are trying to grasp the element of trust within these social networks, the role of blogs cannot be ignored that are not only informative and social but also opinionated. The introduction of Blog track in TREC is a step that will not only boost the performance of opinion detection system but also attract the attention of more researchers. The results of TREC 2006 and TREC 2007 Blog tracks are already shaping the nature of tasks performed on blogs. The thought of making a shift from document level opinion detection to sentence-level opinion detection is also a product of TREC blog track.

Our work in this regard may help researchers to bring their concentrations to a platform specific for challenges that we have discussed. We are seriously thinking to perform this annotation experiment with more than two annotators to make our results more reliable and accurate. Also we are working to develop an evaluation framework for polarity detection on sentence level.

REFERENCES

[1] B. Liu, *Web Data Mining: Exploring Hyperlinks, contents and Using Data, Chapter 11*, Data-Center Systems and Applications, Springer , 2007

[2] G. Mishne and, M. de Rijke, "A Study of Blog Search", Advances in Information Retrieval, Springer, 2006, pp. 289-301

[3] I. Ounis, C. Macdonald, and I. Soboroff, "On the TREC Blog", ICWSM'08: In Proceedings of International Conference on Weblogs

and Social Media, AAAI Press, Seattle USA, March 30-April 2, 2008

[4] M.M.S. Missen, and M. Boughanem, "Sentence-level topic association for opinion detection in blogs", MAW09: In Proceedings of IEEE International Symposium for Web and Data Mining, IEEE Computer Society, Bradford, UK, May 2009, (in press)

[5] M.M.S. Missen, and M. Boughanem, "Using WordNet's Semantic relations for Opinion Detection in Blogs", ECIR 2009: European Conference on Information Retrieval, Springer, Toulouse France, 1-3 April 2009 (in press)

[6] B. Ernsting, W. Weerkamp, and M. de Rijke, "The University of Amsterdam at the TREC 2007 Blog Track", TREC 2007 Blog Track

[7] K. Yang, N. Yu, A. Valerio, and H. Zhang, "WIDIT in TREC-2006 Blog track", TREC 2006 Blog Track

[8] G. Zhou, H. Joshi and C.Bayrak, "Topic Categorization for Relevancy and Opinion Detection", TREC 2007 Blog Track

[9] S. Gerani, M. Carman, and F. Crestani, "Investigating Learning Approaches for Blog Post Opinion Retrieval", ECIR 2009: In Proceedings of European Conference on Information Retrieval, Springer, Toulouse France, 1-3 April 2009 (in press)

[10] T. Wilson, J. Wiebe, and Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", HLT/EMNLP 2005: In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada, October 06 - 08, 2005, pp. 347-354

[11] C. Strapparava, and A. Valitutti, "WordNet-affect: An affective extension of WordNet", LREC 2004: In Proceedings of Language Resources and Evaluation Conference, European Language Resources Association , Lisabon, Portugal, 2004

[12] A. Esuli, and F. Sebastiani, "SentiWordNet: A publicly available lexical resourcefor opinion mining", LREC-06: in Proceedings of Language Resources and Evaluation Conference, European Language Resources Association, Genova, 2006

[13] I. Ounis, M. Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 Blog Track", TREC 2006 Blog Track

[14] C. Macdonald, I. Ounis, and I. Soboroff, "Overview of the TREC-2007 Blog Track"

[15] T. Kiss, and J. Strunk, "Unsupervised Multilingual Sentence Boundary Detection", Computational Linguistics, Vol. 32, No. 4, 2006, pp. 485-525

[16] L. H. Phuong, and H. T. Vinh, "A Maximum Entropy Approach to Sentence Boundary Detection of Vietnamese Texts", RIVF 2008: IEEE International Conference on Research, Innovation and Vision for the Future, IEEE, Vietnam, 2008

[17] M. Stevenson, and R. Gaizauskas, "Experiments on Sentence Boundary Detection", ANLP 2000: In Proceedings of Applied Natural Language Processing Conference, Seattle, Washington, USA, April 29 - May 4 2000, pp. 84-89

[18] R. Blood, *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*, Perseus Publishing, Cambridge, MA, 2002

[19] D. Crystal, Languages and Internet, 2nd Edition Revised, Cambridge University Press, 2006

[20] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions 3rd Edition Chapter: The Measurement of Interrator Agreement*, John Wiley and Sons Inc, 2003

[21] J. Landis and G. G. Koch, "The measurement of observer agreement for categorical data", Biometrics 1977, pp. 159-174

[22] E. Breyten, W. Wouter, and R. Marten, "The Univ. of Amsterdam at TREC 2007 Blog Track, TREC 2007 Blog Track