Sentiment Mining and Indexing in Opinmind

I-Heng Mei
Opinmind, Inc
2900 Gordon Ave, Suite 100-6
Santa Clara, CA 95051
1.408.737.9183
heng.mei@opinmind.com

Hongcheng Mi
Opinmind, Inc
2900 Gordon Ave, Suite 100-6
Santa Clara, CA 95051
1.408.737.9183
charles.mi@opinmind.com

Julius Quiaot
Opinmind, Inc
2900 Gordon Ave, Suite 100-6
Santa Clara, CA 95051
1.408.737.9183
julius.quiaot@opinmind.com

Abstract
This paper presents a production system that efficiently mines social networking sites for sentiments and indexes the expressions for fast retrieval via a web search interface. Sentiment mining is a computational approach used to identify expressions made about topics within a span of text. Social networks represent a particularly rich corpus for mining sentiments because writers express sentiments about a wide variety of topics in their online journals. We introduce a streamlined approach to extract sentiment expressions and target subjects from unstructured and grammatically imperfect text. In addition, we discuss our approach to index sentiment expressions.

1. Introduction
The task of automatically extracting sentiments in blogs has been a focus of study in recent years. There are more than fifty million blogs today and the size of blogosphere is growing daily. Blogs are online journals that allow people to share and publish their thoughts and daily experiences on the web. A typical blog consists of an "about me" section about the author, a series of entries in which the author writes about any topic (s)he chooses, and a series of comments written by readers. An author of a blog can be anyone who is willing to share his or her writings on the web. Similar to traditional written diaries, blogs tend to contain expressions about a wide array of topics. These expressions are often about products, places, and interests though there is no limitation to the topic of expression. Extracting sentiments from blogs can be particularly useful for individuals seeking to understand general sentiments about a topic. Because the sentiments expressed in blogs are unsolicited and voluntary, they also tend to be more genuine and unbiased than sentiments solicited by surveys or focus groups.

The challenges of mining sentiment in blogs stem from the consumer-generated nature of blogs. Firstly, the size and exponential growth rate of the blogosphere require that sentiment mining occur at a sufficiently high speed to ensure reasonable coverage and timeliness. Secondly, blogs are self-published thus blogs more often than not contain grammatically incorrect sentences and fragments. Sentiment mining algorithms must be robust and flexible with ungrammatical and unstructured text. The process of sentiment mining includes identifying sentiment expressions in a span of text and associating sentiment expressions to their targets. Once the sentiment expression-topic sets have been identified, they can be stored for various applications. At Opinmind, we index the sentiment expressions so they can be readily accessed in response to user search queries. Our approach does not rely on a fixed set of target or domain-specific entities so users are able to query for sentiments about any topic they choose.

In this paper, we present an overview of our sentiment mining system and our framework for indexing sentiments. In section 2, we review some related work. In sections 3 and 4, we discuss our sentiment mining approach and the architecture of our sentiment index. Section 5 concludes with a discussion of applications and ongoing work. The system described in this paper is public at www.opinmind.com and is serving more than fifty million opinions from more than five million bloggers.

2. Related work
Hatzivassiloglou and McKeown (1997) first produced a list of seed words to determine whether a sentence contains positive or negative sentiments. Turney (2002) suggested an approach to extend this list by using value phrases composed of six syntactic patterns. Similar to these two approaches, Yi and Nasukawa (2005) built a dictionary of polarity lexicons to extract sentiments from a sentence. Pang and Lee (2004) used SVM to classify subjective sentences by using distance measures between sentences as additional features. Hurst and Nigam (2004) used a document classifier to extract targets of sentiment expressions in a sentence. Traditional machine learning techniques often suffer from domain bias associated with domain-specific training corpuses. And since a sentence has far fewer words or features than a document, running a document classifier on sentences often leads to a very sparse matrix thus significantly reducing recall. The approach described in this paper also operates at the sentence level. However unlike (Yi and Nasukawa, 2005) and (Hurst and Nigam 2004), we employ high precision sentiment activation frames as features to our unsupervised learning algorithm thereby achieving high precision rates while maintaining reasonable recall rates. Our approach doesn't have any bias towards a domain.

3. Sentiment mining
The approach described here consists of two stages – (1) raw text parsing and (2) sentiment extraction. In the first stage, chunks of raw text are broken into sentences and parsed. A heuristic technique is used to determine sentence endpoints. We then parse each sentence into a tree composed of clauses and words with role assignments. Our parser is based on constraints outlined by Blache in Property Grammars (2001). In the second stage, the sentence parse tree is analyzed for sentiments. We reference an extensive lexicon of polarity terms. The polarity lexicon is built by an unsupervised learning algorithm. The unsupervised learning algorithm uses a set of sentiment activation frames as features and continuously learns new polarity lexicons. We have also enhanced the lexicon by adding attributes such as *strength*.

Strength attributes are used to differentiate degrees of expressed sentiments. These additional attributes help us to differentiate between various sentiments and are also learned by our unsupervised learning algorithm.

4. Sentiment indexing

Sentiment and target features are electronically stored in a sentiment index. A sentiment index stores more information about a search term than traditional keyword indexes. Additional information including the polarity of the sentiment, the original sentence, the reference URL and date/time of the entry are stored along with the sentiment and target features in the main index. The main index is segmented and replicated across multiple physical systems to optimize performance and provide redundancy. The run-time system is a distributed network of low-cost commodity computers.

5. Ongoing work and summary

Sarcastic statements are often mis-categorized as it is difficult to identify a consistent set of features to identify sarcasm. One method is to develop a holistic approach which uses previous sentiments expressed by the author to determine the probability of a sarcastic sentiment. In addition, topics of interest are often continued from previous conversations. To precisely identify these topics, one method is to develop a history of mentioned topics and associate them with the appropriate sentiment expression.

Work at Opinmind to date has focused on optimizing the speed and accuracy of sentiment mining and building a scalable and efficient sentiment index. We aim to continually improve our accuracy and recall in order to fulfill our goal of collecting and presenting sentiments of bloggers all over the world.

6. References

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the ACL, 2004*.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 105-112.

Blache P. & J.-M. Balfourier (2001) "Property Grammars: a Flexible Constraint-Based Approach to Parsing", In *Proceedings of IWPT-2001*.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania.

Hongcheng Mi, I-Heng Mei. Searching Sentiments in Blogs. In *American Association for Artificial Intelligence 2006.*

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, pages 174-181*, Madrid, Spain, July. Association for Computational Linguistics.

Kamal Nigam, Matthew Hurst. 2004. "Towards a Robust Metric of Opinion." In *American Association for Artificial Intelligence 2004*.

Jeonghee Yi, Tetsuya Nasukawa. 2004. "Sentiment Analysis: Capturing Favorability Using Natural Language Processing" In *K-CAP 03, October 23-25, Florida, USA.*