The University of Birmingham
School of Computer Science

Research Student Monitoring Group

RSMG 1

Name of student: Phillip Smith

Start date: 4-Oct-2010

Registration number: 781797

Previous degrees: Computer Science BSc. Hons. (Class I)

Name of supervisor: Dr. Mark Lee

Date of report: 11-Nov-2010

Proposed working title: Combining Classification Techniques to Improve Sentiment Analysis

Sentiment analysis is the process of analysing affective phrases within the unstructured text of a document, so that an emotion may be automatically associated with it. A current approach to sentiment analysis is to perform this task at a coarse-grained level (Liu, 2010), categorizing text into two overarching categories of how positive or negative a term is. This is adequate in situations where a shallow analysis is required, such as evaluating online product reviews. However, to gain an understanding of the emotional content that is being communicated in a document, sentiment analysis should be performed using a granular model of emotion. Determining whether it is possible to formalize a computational model of sentiment that encompasses a comprehensive model of emotion, is a primary motivation of this research.

To formalize the problem, we must take into account the issues that natural language processing (NLP) tasks hold, and observe the methodologies that have been employed to solve them. Sentiment analysis can be generalised as a text classification problem. Approaches to this problem can be viewed in three ways: as either a supervised (Yessenalina et al., 2010), an unsupervised (Turney, 2002), or a semi-supervised (Sindhwani & Melville, 2008) attempt at accurately determining the sentiment of a document. Supervised approaches, such as using decision trees or support vector machines, attempt to fit a predetermined model of sentiment to the text that it will be applied to. The key to this approach is having a suitably robust model to be utilised initially. Using an opposing methodology, unsupervised algorithms aim to seek out patterns in the input document set, through use of data clustering techniques. From analysing the output, a model of emotion suggested by the resulting clusters can be formalized. Semi-supervised algorithms make use of the previous two approaches; they take a mixture of both labelled and unlabeled data as its training set in order to generate an appropriate model, and evaluate a data set.

This research aims to analyse the above techniques, to determine which combination of procedures yield the highest accuracy in a domain independent environment. It can be hypothesised that a single method alone will not produce optimal results, due to the difficulties associated with classifying text into appropriate classes. It is therefore believed, that combining the aforementioned classification techniques, in addition to establishing a well-defined computational model of sentiment, are all necessary steps in improving sentiment analysis.

Bibliography
Liu, B. (2010). 'Sentiment Analysis: A Multi-Faceted Problem.',
http://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf, [accessed 11 Nov 2010].

Yessenalina, A. & Choi, Y. & Cardie, C. (2010), 'Automatically generating annotator rationales to improve sentiment classification', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics(ACL '10)*, pp.1386–1395.

Turney, P. (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 (ACL '02)*, pp. 417-424.

Sindhwani, V. & Melville, P. (2008) , 'Document-Word Co-regularization for Semi-supervised Sentiment Analysis', in *Eighth IEEE International Conference on Data Mining (ICDM '08)*., pp.1025-1030.

Proposed members of the thesis group:

1 Student: Phillip Smith
2 Supervisor: Dr. Mark Lee
1. RSMG member: Dr. Peter Hancox
2. Non-RSMG member: Prof. John Barnden

Has the Ethical Self-Assessment Form been completed?
Was further ethical review needed, and if so has it been applied for?

(Ethical review needs to be completed by the Thesis Proposal stage at the end of year 1, but there is no harm in doing it early.)

Any comments by the student or supervisor:

Signed (Supervisor):

Signed (Student):

Date:

Please return this form to the Research Students Tutor by 12 November, 2010.
You must also attach a skills development form **GRS1A**.