Processing Sentiments and Opinions in Text

NG, Jun Ping
National University of Singapore
ngjp@nus.edu.sg

Abstract

There is a recent surge in interest in processing sentiments and opinions in free text. I examine research in this area, and summarise the differing approaches to the various sub-tasks and applications relating to sentiment analysis. A list of existing evaluation efforts for sentiment analysis tasks is also presented. Finally, 2 possible research directions that could help to further the state-of-the-art are suggested.

1 Introduction

1.1 Problem

Much work has gone into the processing and understanding of objective expressions of factual text. In recent years however researchers have started to move towards processing subjective sentiments and opinions expressed in natural language. This area of research is so new that the field has till date yet to agree on a set of terminology to describe it. I will follow the approach adopted by (Pang and Lee, 2008) and refer to this new area of study as either *sentiment analysis* or *opinion mining* interchangeably.

Borrowing a definition from (Liu, to be published in 2010), sentiment analysis (or opinion mining) is the computational study of opinions, sentiments and emotions expressed in text. (Esuli and Sebastiani, 2005) identifies 3 specific subtasks that make up sentiment analysis :

1. *Determining subjectivity*
   Involves deciding whether a piece of text is factual or subjective (ie. expresses an opinion about a particular topic)

2. *Determining polarity*
   Involves deciding whether the expressed sentiment of a subjective piece of text is positive or negative

3. *Determining strength of polarity*
   Involves grading the intensity of the expressed sentiment in a subjective piece of text

1.2 Application

Sentiment analysis is useful for a wide-range of scenarios. Consider a review of a flute (referred to by its manufacturer *Muramatsu* and name *EX*) (Rev, 2008)[1] :

> (1) Body and key work on the Muramatsu EX is well made. (2) But for me, the sound from the head is awesome! (3) I play flute reasonably well and with significant experience, and I service many different flutes. (4) This and the GX are the ones I really look forward to play testing. (5) (There are a few others that show up too, but more pricey.) (6) Excellent value for money. (7) Would like to try that head on a good cheap body such as Yamaha 200 series.

The manufacturer Muramatsu may like to find out whether there are more positive reviews of its flutes than negative ones in general. Given a set of documents, **document classification**[2] identifies whether the text in each document express a positive or negative opinion. This can also be performed in a more fine-grained manner — **Subjectivity classification** decides whether sentences within each document carries a subjective opinion, and if so whether the opinion is positive or negative.

A flute player looking to purchase a new instrument on the other hand will be more interested to find out about specific qualities of the flute, such as it's reliability or sound. **Feature aggregation** involves identifying features of the topic described within a piece of text, and associating it to a subjective opinion (for example from (2) in our Muramatsu example — *sound* $\mapsto$ *awesome*).

Further, the flute purchaser may also want to know how the flute compares with another. **Com-**

[1] Sentences are labelled to make it easier to reference them subsequently.

[2] Common literature also refers to this as *sentiment classification*

parative sentiment analysis[3] looks at *comparative sentences* and ranks entities based on common features (for example from (5) in our Muramatsu example —*Price* ↦ *cheaper* ↦ *Muramatsu EX*).

1.3 Difficulties, Challenges and Issues

It is non-trivial to solve the 3 sub-tasks identified by (Esuli and Sebastiani, 2005). Some of the obstacles and problems that need to be overcome are briefly explained in the sub-sections that follow.

1.3.1 Keywords are Insufficient

At first glance, document and subjectivity classification may lend themselves to a traditional bag-of-words approach, where the presence of certain keywords can help decide on the expressed sentiments. Generally this is true as text with terms like *good*, *fantastic* and *excellent* typically is more positive than another with terms like *lousy*, *unreliable* and *terrible*. However there are too many ways through which sentiment can be expressed. The inclusion of the word *not* for instance changes the opinion associated with *good* totally. Sentiment can also be defined without the use of any of the above "opinion words" such as in this review of a fragrance quoted from (Pang and Lee, 2008) :

> If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.

1.3.2 Facts or Opinions?

It is often hard to decide whether a text states an objective fact, or expresses an opinion. Returning to our example on the Muramatsu flute, sentence (5) is a fact about the relative prices of different flutes, but yet the follow up sentence (6) though relating to the same idea of affordability is an opinion about the suitability of the pricing.

1.3.3 Context

The context in which a piece of text is found also plays a large part in determining the sentiment of the text. A sentence such as *Go read the book.* is likely a positive one when it is part of a book review, but less so when it is extracted from a movie review. It is a challenge to decipher and put to use the context of a piece of text, and this is still an open problem till date.

1.3.4 Spam

Another potential problem for sentiment analysis is sieving through *opinion spam*. The rise of WEB 2.0 with its emphasis on user-generated content has brought about a huge increase in the amount of opinionated text on the Internet, yet it is unclear how many of these are untruthful, or overly biased. Opinions could have been fabricated by unscrupulous manufacturers for example, or be unfairly influenced by over-zealous contributors. Often just looking at the text alone is not sufficient to filter such spam and even humans may find it hard to discern between genuine feedback and fabricated ones.

1.4 Related Work

There has already been some literature exploring the topic of sentiment analysis. (Pang and Lee, 2008) presented for example a thorough survey on sentiment analysis, focusing on explaining modern statistical approaches to extracting as well as aggregating sentiments from documents. Work by (Liu, to be published in 2010) attempts to define formally the different sub-tasks of sentiment analysis, as well as visit some of the issues involved in the various sub-tasks. Complementing these two surveys, this paper aims to present an updated overview of the state-of-the-art of approaches to sentiment analysis, and highlight their strengths and weaknesses with respect to the issues raised earlier.

1.5 Organisation

In the next section, a survey of research work in the field is presented. The subsequent section then explores various performance measures and benchmarks against which work in sentiment analysis is evaluated. A look at possible future extensions and directions of work in this area then rounds up this paper.

2 Current Approaches

Research in the field is typically targeted at the various sub-tasks of sentiment analysis. The survey of the various pieces of work will thus be taxonomised according to the sub-task they are intended to solve. It is worth noting here however that solutions that determines the polarity of a piece of text will also generally determine the

[3] There is again no widely adopted terminology for this task. (Pang and Lee, 2008) refers to this as *sentiment analysis of comparative sentences*, while (Ganapathibhotla and Liu, 2008) terms this as *opinion mining of comparative sentences* for example.

subjectivity of the piece of text. Work on these 2 sub-tasks will as a consequence be discussed together.

There has also been much work looking beyond solving the basic sub-tasks of sentiment analysis. Two other active areas of research include (1) feature aggregation, and (2) analysis of comparative sentences. Feature aggregation looks at identifying specific *object features* that are commented on, and then determining whether the opinions of these features are positive or negative. For example in sentence (1) of the flute review quoted earlier, *production quality* is a likely feature of interest, and it is reviewed to be positive. Analysis of comparative sentences studies opinions expressed in the form of comparisons. One such comparison can be *The sound of the Muramatsu EX is better than the Yamaha YF-221.* These work are crucial to the application of sentiment analysis to product reviews and summarization among others, and thus are reviewed here too.

2.1 Subjectivity and Polarity

2.1.1 Machine Learning

Work in determining the subjectivity and polarity of text falls into 1 of 2 possible categories most of the time. One important thread of work in determining subjectivity and polarity is the use of machine learning. Machine learning methods are generally either (1) supervised, or (2) unsupervised.

2.1.1.1 Supervised Learning

Supervised learning approaches make use of large corpora of annotated text and state-of-the-art text classifiers for subjectivity/polarity classification. The various approaches differ in the complexity of the features used. These features range from just making use of simple lexical information, to complex ones making use of discourse structures and analysis.

At one end of the spectrum are features making use of lexical information. Features including n-grams, word frequency, and part-of-speech tags are used with classifiers such as the naive Bayes classifier, maximum entropy classifier and support vector machines (Pang et al., 2002; Dave et al., 2003). A classifier based on these features can also be combined with a rule-based approach as was done in (Prabowo and Thelwall, 2009) for better results. In this work, rules reasoning about keywords such as $\texttt{more} \wedge \texttt{expensive} \mapsto +$ are combined with a statistical classifier.

Looking beyond local textual features, it is often useful to also consider the topic a piece of text is describing. The intuition for this is that not every opinionated sentence in a document describes a sentiment relevant to a target topic. One way to integrate such topic considerations is to make use of a generic text classifier like the Winnow classifier (Hurst and Nigam, 2004) to first decide if a document is relevant to the target topic. If it is, the individual sentences that make up the document are then put through the classifier again to verify their relevancy. Sentences which match the topic at hand are then analysed for subjectivity expression. An alternative approach is to learn a joint probabilistic model of the topic discussed in the text, along with the expressed sentiments like in (Mei et al., 2007).

(Esuli and Sebastiani, 2005) combines different resources including WORDNET, dictionaries and thesauruses by first expanding representative anchor words. The gloss of the anchor words and their expanded synonyms are next used in a bag-of-words fashion to train a classifier to decide polarity.

More recently, a fair amount of work has gone into making use of discourse related features for better supervised learning. Text snippets for example may *reinforce* each other, as highlighted by sentences (5) and (6) in the running Muramatsu example. Both sentence attest that the flute is fairly priced. Sentences (1) and (2) however are *non-reinforcing* as they do not express similar stances. Work by (Somasundaran et al., 2009) has shown that these discourse features are beneficial in deciding on the polarity of a piece of text.

Besides analysing the discourse relation between sentences, a more coarse-grained approach can also be taken by tracking the relationship between segments of a document. The postulate for these approaches is that a degree of continuity exists between text segments, and thus neighbouring segments should receive similar polarity labels (Pang and Lee, 2004; Thomas et al., 2006).

2.1.1.2 Unsupervised Learning

As annotated data may not always be easily available, many researchers have also worked on unsupervised approaches. One of the earliest work in this area deals with the creation of a lexicon of *sentiment* or *opinion* words to support a bag-of-words approach to sentiment classification. (Hatzivassiloglou and Wiebe, 2000) makes use of a log-linear model to identify regularly inflected adjectives, which are then used to predict the subjectivity of a sentence. The intuition for the approach is that such adjectives often express properties in varying degrees of strength and can be a suitable predictor of the subjectivity of sentences.

Another line of work by (Turney, 2002) determines the subjectivity of words by computing the mutual information between each word and an anchor word like *excellent* or *poor*. Words which are similar to the anchor words are highly likely to be used in a subjective context, and can therefore be used to judge the opinion expressed in a sentence.

Besides use of lexicons to predict subjectivity and polarity, a bootstrapping approach can be adopted to learn a classifier for sentiment classification. Typically such a process involves using (1) a simple classifier to identify seed annotated data, and (2) use the seed data to iteratively improve on the classifier. (Riloff and Wiebe, 2003) for example makes initial use of a classifier to automatically annotate a seed dataset. The classifier identifies subjective sentences by using simple methods including a lexicon-based bag-of-word feature. From this seed dataset, word patterns stemming from a pre-defined set of templates (for eg. `Subject` $\mapsto$ `Passive-Verb`) are learnt and associated with the subjectivity (or the lack of) of sentences.

2.1.2 Making Use Of Graphs

Similar to the work of (Turney, 2002) described above, where subjectivity is determined by the mutual information between words and an anchor word, (Kamps et al., 2004) describes another approach to deciding subjectivity with a distance metric. A graph is built from important anchor words (like *good*) and their synset are extended by recursively expanding each new word and corresponding synset. The distance metric is computed as the number of nodes between two different words, normalised by the distance of two opposite anchor words like *good* and *bad*. A word nearer to either of the anchor words will be classified accordingly.

2.2 Strength of Polarity

Besides subjectivity and polarity classification, determining the strength of polarity is the third sub-task of sentiment analysis. In this paper, I will be referring to this also as strength classification.

(Wilson et al., 2004) made use of a supervised learning approach for strength classification. Their work combined well-known features, including the use of keywords, adjectives annotated for subjectivity, and n-grams, with syntactic features identified from both constituent grammar and dependency parses. These features were used then with 3 different learning algorithms including boosting, rule-learning and support vector regression.

2.3 Feature Aggregation

Having gone through work on the various sub-tasks of sentiment analysis, this and the next section explores work in applications of sentiment analysis.

The same methods used for the earlier described sub-tasks of sentiment analysis can be used for feature aggregation. (Liu et al., 2005) found that a supervised learning approach using features such as part-of-speech tags, n-grams, and stemming is useful for identifying and extracting product features from reviews. They also suggested a visualisation framework which allows users to quickly compare opinions of two different products.

In (Popescu and Etzioni, 2007), noun phrases are initially identified as potential features. The pointwise mutual information score between each phrase and meronymy[4] discriminators associated with the target product is calculated and subsequently used to filter out phrases which are not features.

Augmenting these automatic approaches, (Wei et al., 2009) proposed a semantic-based refinement where adjectives with word senses that are used in an objective context are removed from a list of opinion words used for subjectivity classification. The motivation behind this approach is that the multiple word-senses of such words may lead to more false positives.

2.4 Analysis of Comparative Sentences

Sentiment analysis of comparative sentences can typically be solved in 2 difference steps : (1) iden-

[4] A meronym denotes a constituent part of, or a member of something. For example, 'finger' is a meronym of 'hand' because a finger is part of a hand. (Wik, 2000)

tifying comparative sentences, and (2) extraction of object and object features.

A bag-of-words approach has been shown (Jindal and Liu, 2006a) to work reasonably well in identifying comparative sentences. It is observed that most comparative sentences generally consists of keywords or key phrases denoting comparisons.

Separately (Bos and Nissim, 2006) makes use of a combinatory categorial grammar (CCG) parser to interpret and identify superlatives within sentences. The use of the CCG parser makes it possible to make use of part-of-speech tags in tandem with CCG categories to recognise superlative adjectives. Having identified the comparative sentences, information extraction techniques can then be applied much as in feature aggregation to identify objects and their features. (Jindal and Liu, 2006b) for example makes use of a rule-based system together with conditional random fields (CRF) to extract the this information.

3 Benchmarking and Evaluation

Work in sentiment analysis is increasingly gaining momentum. As new research tasks and applications emerge, there is a need to be able to systematically evaluate these work objectively. In this section I will take a look at some of the efforts targeted at achieving this.

3.1 Shared Tasks

Shared tasks in workshops and conferences provide a common dataset against which participating systems can be evaluated. There has been several such tasks concerning sentiment analysis.

The Multilingual Opinion Analysis Task (MOAT) 2008 involves many of the sentiment analysis sub-tasks. These include for example subjectivity classification, polarity classification, strength classification, and detection of opinion holders and targets. Plans are in place to expand MOAT 2009 to include cross-lingual opinion analysis, where a query on opinions needs to be answered from source text of other languages.

The Content Analysis for the Web (CAW) 2.0 workshop in the International Conference on the World Wide Web (Codina et al., 2009) organised a sentiment and opinion mining shared task. The shared task comprises of 2 tracks : (1) sentiment analysis, and (2) opinion analysis. Sentiment analysis in this case refers to associating given text to 1 of 4 emotional categories (ie. neutral, happy, angry, and sad), while opinion analysis refers to associating given text to 1 of 3 subjectivity categories (ie. factual, opinionated-positive, opinionated-negative). Training data for the shared task includes text crawled from popular Web 2.0 sites including TWITTER, MYSPACE, and SLASHDOT.

In the upcoming SemEval-2 in 2010, there will be a task looking to disambiguate sentiment ambiguous adjectives. The tasks stems from the observation that the subjectivity of adjectives can be context-dependent. For example the adjective *high* in the snippets *the price is high* and *the quality is high* expresses opposite polarities.

Besides the above tasks focusing on specific sub-tasks of sentiment analysis, there are also efforts to evaluate applications of sentiment analysis. The Text Analysis Conference (TAC) in 2008 for example ran an opinion question-answering (QA) track, which requires systems to locate answers to opinion questions. Opinion questions are questions which seek to find out the expressed sentiment of a piece of text, such as *What do people like about Ikea?*. The summarisation track in the same TAC 2008 also included an opinion summarisation pilot task, where the goal is to write summaries of opinions from blogs.

3.2 Data sets

Annotated data sets are an important part of machine learning approaches to sentiment analysis. Common data sets also provide a good starting point from which systems can benchmark themselves to. A good survey of the various data sets available is done in (Pang and Lee, 2008).

Available data sets can generally be differentiated based on the sources from which they are derived. Unstructured data sets tend to be composed of free-styled content posted by web users. For example the *Blog06* corpus consists of blog posts crawled from the web. Structured data sets on the other hand are typically from specific opinion sources such as reviews from e-commerce sites or interest sites. An example is the *Cornell movie-review* data set (Pang et al., 2002) which is assembled from a newsgroup of movie reviews.

4 Future work

Much of the work reviewed in this paper attempts to make use of well-known techniques like machine learning to solve the new tasks associated with sentiment analysis. Though important, work in feature engineering has yet to yield compelling innovations. As work in this emerging fields mature, it is likely that the results from existing approaches will plateau.

4.1 Fusion of Approaches

One way to improve on system efficacy can be to combine the various approaches and weld them together for better performance. Work targeting lexicons, syntax, and discourse individually have been reviewed. If these approaches are combined for an unified approach to sentiment analysis, it is highly likely that better performance can be obtained.

It is possible however that combining all the various features together will lead to a problem of data sparsity. The solution around this may be to note that the influence each feature exerts for different text input is not likely to be a constant. Clues from the lexicon used and construction of input text may be used to dynamically adjust the feature space, selectively enabling/disabling various features. This may be very useful to help circumvent the curse of dimensionality.

4.2 Deep Semantics

Conspicuously missing from the work reviewed here are deep semantic approaches. There has been attempts to make use of semantic information through the use of keyword-aware rules (Wei et al., 2009) or logical representations (Bos and Nissim, 2006). However such work has yet to be embraced by the community at large.

The successes of knowledge-rich approaches in areas like question-answering (Moldovan et al., 2007) and recognising textual-entailment (Raina et al., 2005) hint that it may be worth the while to explore such approaches for sentiment analysis. In particular, such approaches should serve as a good complement to their knowledge-poor counterparts.

In fact, recall the review of a fragrance listed earlier in Section 1.3.1 where an opinion is expressed without the use of any opinion word. It is unlikely that a knowledge-poor approach will be able to identify the implicit scorn and sarcasm in the text snippet. While current advances in semantic-based approaches are also yet able to make sense of this example, this line of work should hold more promise in extracting such embedded sentiments.

5 Conclusion

In this work, I have reviewed the tasks relating to processing sentiments and opinions in text. Major approaches to the various sub-tasks including document and subjectivity classification, polarity detection and strength classification are examined, along with important applications like feature aggregation and the sentiment analysis of comparative sentences. Some of the recent evaluation exercises concerning sentiment analysis tasks are also explained. Lastly, a discussion on possible extensions to current research is presented.

References

Johan Bos and Malvina Nissim. 2006. An empirical approach to the interpretation of superlatives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Joan Codina, Andreas Kaltenbrunner, Jens Grivolla, Rafael E. Banchs, and Ricardo Baeza-Yates. 2009. Content analysis in web 2.0. In *Proceedings of the International World Wide Web Conference*, April.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In *Proceedings of the International Conference on World Wide Web*.

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 617–624.

Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences.

V. Hatzivassiloglou and J.M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the Conference on Computational Linguistics*, pages 299–305.

Matthew Hurst and Kamal Nigam. 2004. Retrieving topical sentiments from online document collections. *Document Recognition and Retrieval*, XI:27–34.

Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*.

Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relati. In *Proceedings of the National Conference on Artificial Intelligence*.

J. Kamps, M. Marx, R.J. Mokken, and M. De Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of the International Conference on Language Resources and Evaluation*, volume 4, pages 1115–1118.

B. Liu, M. Hu, and J. Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the International Conference on World Wide Web*, pages 342–351. ACM New York, NY, USA.

Bing Liu, to be published in 2010. *Handbook of Natural Language Processing*, chapter Sentiment Analysis and Subjectivity.

Q. Mei, X. Ling, M. Wondra, H. Su, and C.X. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the international conference on World Wide Web*, pages 171–180.

Dan Moldovan, Christine Clark, and Mitchell Bowden. 2007. Lymba's poweranswer 4 in trec 2007. In *Proceedings of the Text Retrieval Conference*.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity sumarization based on minimum cuts. In *Proceedings of the Association for Computational Linguistics*, pages 271–278.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, July.

Ana-Maria Popescu and Orena Etzioni, 2007. *Natural Language Processing and Text Mining*, chapter Extracting Product Features and Opinions from Reviews, pages 9–28. Springer.

Rudy Prabowo and Mike Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3:143–157.

Rajat Raina, Aria Haghighi, Christopher Cox, Jenny Finkel, Jeff Michels, Kristina Toutanova, Bill MacCartney, Marie-Catherine de Marneffe, Christopher D. Manning, and Andrew Y. Ng. 2005. Robust textual inference using diverse knowledge sources. In *Proceedings of the PASCAL Recognising Textual Entailment Challenge*.

2008. Review centre. `http://www.reviewcentre.com/reviews62286.html`.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, July.

Chih-Ping Wei, Yen-Ming Chen, Chin-Sheng Yang, and Christopher C. Yang. 2009. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management*.

2000. Wikipedia – the free encyclopedia. `http://en.wikipedia.org/`.

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.