International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 3, Issue 1
April 2014

Opinion Mining, Analysis and its Challenges

Nidhi R. Sharma , Prof. Vidya D. Chitre
Computer Department, University of Mumbai, Computer Department, University of Mumbai,
[1]nidhipranjan@gmail.com
[2]vidyadchitre2k14@gmail.com
BVCOE, Navi Mumbai

***Abstract*--Any user , buyer or customer rely on the Web for their opinions on various products and services which they have used, it is very important to develop methods to automatically classify and evaluate them. The task of classifying and analyzing such collective data together is known as customer feedback or review data, and is called as opinion mining.**

Opinion Mining is a very challenging and promising discipline which is defined as an intersection of information retrieval and computational linguistic techniques to deal with the opinions expressed in a document. The main aim at solving the problems related to opinions about products, reviews ranking in movies, Politian in newsgroup posts, review sites etc. In this paper we are about to cover the source of data from where we take , its classification, evaluation process and then grouping techniques, tools used, and future challenges in opinion mining. Opinion mining consists of various stages such as extraction of data from various sources, text classification, grouping together and then evaluating it to positive or negative or true or false value. On the basis of our survey and analysis of the techniques, we provide an overall picture of what is involved in developing a software system for opinion mining.

Keywords-- Data mining, web mining, opinion mining, sentiment classification, text classification, evaluations.

1. Introduction:

The World Wide Web is growing at an alarming rate not only in size but also in the types of services and contents provided. Each and every users are participating more actively and are generating vast amount of new data. These new Web contents include customer reviews and blogs that express opinions on products and services – which are collectively referred to as customer feedback data on the Web. As customer feedback on the Web influences other customer's decisions, these feedbacks have become an important source of information for businesses to take into account when developing marketing and product development plans. This era is of automated systems [1] and digital information every field of life is evolving rapidly and generating data. As a result huge amount of data produce in the field of science, engineering, medical, marketing, finance etc [2]. Automated systems are needed to automate analysis, summarization, and classification of data. It also helps at enterprise level to take related decisions. Multiple research fields like statistics, machine learning, artificial intelligence and visualization are involved to develop such automated systems [2-7].

A number of proficient ways are existing [4] to store the huge volumes of data, computational techniques and models are required to extract the hidden patterns and knowledge. These techniques and tools are used to transform the data into useful information, to make market analysis, fraud detection and find the customer intentions etc. Such computational tools and techniques are the subject of *Knowledge Discovery in Database and Data Mining* [4-5].Text mining is an interdisciplinary method used in different fields like machine learning, information retrieval, statistics, computational linguistic and data mining to form mining algorithms. Some researchers defined text mining as tool to discover the new knowledge from huge volume of natural language text using computational algorithms. Web mining is a sub discipline of text mining used to mine the semi structured web data in form of web content mining, web usage mining and wed structure mining.

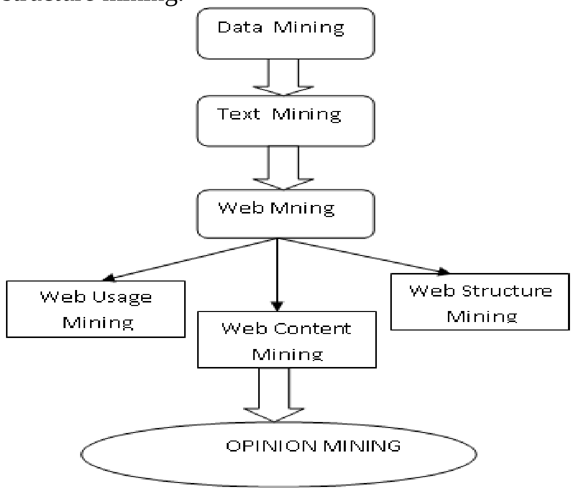

Fig. 1. Data Mining Hierarchical Model

This paper is organized as follows: section 2 covers opinion mining. Section 3 is about the dataset source. Section 4 is about the levels of sentiment classification. Section 5 describes text classification Section 6 is all about the grouping feature Section 7 describes evaluation process and Section 8 describes various recent tools used to do this.

2. Opinion Mining (O.M)

Opinion Mining is a promising discipline which is defined as combination of information retrieval and computational linguistic techniques deals with the opinions expressed in a document. The field major goals is solving the problems related to opinions about products, politics in newsgroup posts, review sites, etc. There are different techniques for summarizing customer reviews like Data Mining, Information Retrieval, Text Classification and Text Summarization [2], before World Wide Web users asked the opinions of his family and friends to purchase the product . In the very same way when any organizations need to take the decision about their products they had to conduct various surveys to the focused groups or they had to hire the external consultants to do so [4]. Web 2.0 [7], ease the customers to take decision to purchase the product by reviewing the posted comments. Customers can post reviews on web communities, discussion forums, twitters, blogs, product's web site these comments are called user generated contents. Web2.0 is playing a vital role in data extracting source in opinion mining. It facilitates users to know about the product from other customer's reviews who have already used it instead of asking friends and families. Companies, instead of conducting surveys and hiring the external consultants to know about the clients opinions, extract opinionated text from product web site [8]. An automated opinion summarization model is needed to complete these tasks. Opinion Mining or Sentiment Analysis is the area to extract the opinionated text datasets and summarize in understandable form for end user [8]. Opinion mining is used to extract the positive, negative or neutral opinion summary from unstructured data. It involves subjectivity in text and computational management of opinion. It is the sub-discipline of web content mining, which involves Natural Language Processing and opinion extraction task to find out the polarity of any product consumers feedback [4]. Figure 2 describes the object model of Opinion Mining.

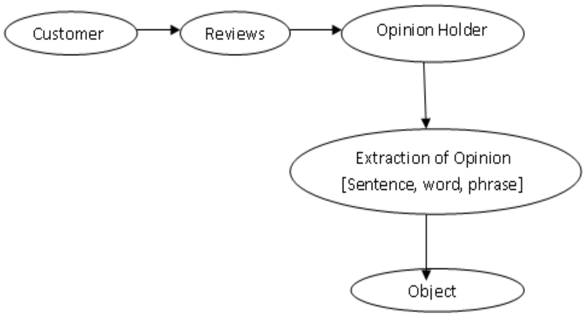

Fig. 2. Opinion Mining Model

In the above diagram there are five components i.e a customer giving the broader reviews from various sources , it is the sentiment, views or judgment about any object based on knowledge or experience , then is the Opinion Holder which is the person, organization that expresses its views or sentiments about any object and the Object which is an entity (person, topic, product or organization) about which the opinion expressed.

3. Data Source

People and companies across disciplines exploit the rich and unique source of data for varied purposes. The major decisive factor for the improvement of the quality services rendered and enrichment of deliverables are the user consumers opinions. Review sites, blogs and micro blogs provide a good understanding of the reception level of products and services.

3.1 Review Sites

Opinions are the major and actual data or more precise a decision for any user in making a purchase. The user generated reviews for products and services are mainly available on internet. The sentiment classification uses reviewer's data are gathered and composed from the websites likewww.gsmarena.com (mobile reviews),www.amazon.com (product reviews), www.CNETdownload.com (product reviews), which hosts millions of product reviews by consumers. [9]

3.2 Blogs

The name associated to universe of all the blog sites is called blogosphere. People write about the topics they want to share with others on a blog. Blogging is a happening thing because of its ease and simplicity of creating blog posts, its free form and unedited nature. We find a large number of posts on virtually every topic of interest on blogosphere. Sources of opinion in many of the studies related to sentiment analysis, blogs are used.[10]

3.3. Micro-blogging

A very accepted communication tool among Internet users is micro-blogging. We use this as one of the data source formed as a dataset of collected messages from Twitter. Twitter contains a very large number of very short messages created by the users, consumers of this micro blogging platform. Millions of messages appear daily in well-liked web-sites for micro-blogging such as Twitter, Tumblr, Facebook. Twitter messages sometimes express opinions which are used as data source for classifying sentiment. [11]

4. Sentiment Classification

4.1 Document level:

Document level sentiment classification is based on the sentiments executed on the overall sentiments expressed by authors. Documents classified according to the sentiments instead of topic. It is very useful in summarizing the whole document as positive or negative polarity about any object (camera, fridge, mobile, car, movie, and politician). In [12] authors proposed a new approach “classification of opinion documents by a vote system” based on combining text representations using key-words related to bigrams. Sentiment Classification Using Phrase Patterns in used Special tags opinion words. System constructed some phrase patterns and compute sentiment orientation using unsupervised learning algorithm. Proposed system achieved 86% accuracy. Investigated perspective from which a document was written. They build Naïve Bayes based model and test on Israeli-Palestinian conflict. Their corpus consists of articles published on the bitter lemons website. They used NB-B (full Bayesian inference) and NB-M (Maximum a posteriori).

4.2 Sentence level:

Sentence level sentiment classification models is used for the extraction of the sentences contained in the opinionated terms, opinion holder and opinionated object. It is one level deep to document level and just concerns to the opinionated words but not the features. Total number of positive and negative words are counted from the extracted and classified sentences and if positive words are maximum then opinion about object is positive and if the negative words are more than opinion object is negative otherwise the opinion object will be neutral. To mine the customer reviews on a product proposed unsupervised algorithm is used and in this the algorithm find frequent features using Apriori algorithm. Chinese WordNet set classify opinion words in clauses (pos, neg or neutral) to summarize the comments. Sentence level opinion mining uses subjective and polarity (orientation) to find strength of opinions at the clause level. [13], all these are a notable work in this regard. To find the strength of opinions a new idea of syntactic clues is used. They use a wide range of features to find the strength of opinions. The system is about to provide tools and support for information analysts in government, commercial, and political domains, who want to be able to automatically track attitudes and feelings in the news and on-line forums. Opinion Analysis based on Lexical Clues and their Expansion to improve combination of rule-based algorithms and machine learning techniques. [14] proposed semi-supervised learning method based on highly precise seed rules. Subjectivity discovered at the sentence level. Polarity of the sentence defined as Positive, Negative or Neutral as well as opinion holders identified. The experimental results demonstrate that system achieved 45% Accuracy to extract opinionated sentences and 35% Accuracy to identify opinion holders.

4.3 Feature based level:

In customer reviews document, reviewer express positive, negative or both sentiments about the object and attributes. Document level and sentence level classification does not tell the likes and dislikes of consumer about particular attributes of object [16]. When consumer comment on object (product, person, and topic, organization) he comment on the features of object[15].For example, if users commented on a Mobile Phone they basically comment on Camera result, LCD size, speaker, weight etc. On camera output 125 comments express the positive opinions and 25 comments may be negative. If a new customer is interested in camera quality of mobile he can take decision easily to purchase the product or not. To explore the detailed opinion on product or any topic, a detailed opinion mining study is required that is called feature based opinion mining [17]. Statistical Opinion Analyzer (SOA) extract the polarity of online customer reviews using Bayesian probability and frequency distribution. The proposed system helps the new customer to purchase the product and manufacturer to enhance the product's functionality. Reviews crawled, preprocess, tagged (GO tagger) and insert in SOA to find the positive and negative opinion probability and frequency distribution as well. The proposed system originated the very promising results. In a web based system SUMView crawled reviews from Amazone.com, decompose into sentences and tagged to find the nouns and noun phrases. Product features extracted using Hu and Liu (2004) method and top five extracted features were suggested to the users on the basis of frequency.

Here we describes the shortcomings of sentiment classification at documents, sentence and feature level.

Document level classification

- It does not give details of what people likes or dislikes because writer comments only on the specific aspects of product.
- It is also not applicable on forums and blogs as they contains only few opinionated sentences on features of object.
- It defines the polarity of document , but a positive phrase does not indicates that the user likes everything and similarly a negative phrase does not indicate that the opinion holder dislikes everything.

Sentence level classification

- It is a fine-grained level of classification in which polarity of the sentence can be given by three categories as positive, negative and neutral.
- In this the identification features indicate whether sentences are on-topic which is kind of co-reference problem.

Feature level classification

- It is defined as product attributes or components.
- In this approach positive or negative opinion is identified from the already extracted features.
- It is a fine grained analysis model among all other models.
- It is having a drawback that it could really cut very badly if there used any grammatically incorrect text.

5. Text Classification

Now a day's huge and vast structured and unstructured volume of online text is available through different websites, internet news feed, emails, cooperate databases and digital library. The main problem is to classify text documents from such massive databases. Using set of training labeled examples statistical text learning algorithms can be trained to approximately classify documents. The news articles and web pages were automatically catalogued by these text classification algorithms.

Naïve Bayes Classifier is a well known probabilistic classifier which describes its application to text. In order to incorporate unlabelled data, the foundation Naïve Bayes was build. The task of learning of a generative model is to estimate the parameters using labeled training data only. The estimated parameters are used by the algorithm to classify new documents by calculating which class the generated the given document belongs to. Biological data is modeled using probabilistic models, such as HMM (Hidden Markov Model) or Bayesian networks which are efficient and robust procedures for learning parameters from observations. There are various sources for missing values such as in medical diagnosis, missing data for certain tests or gene expression clustering due to intentional omission of gene-to-cluster assignments in the probabilistic model. Such error does not occur in EM algorithm.[18,19]

6. Grouping Features

Regular expressions: Regular expressions are highly specialized programming language, in which the rules are specified for the set of possible strings that can be matched and the set might contain English sentences, or e-mail addresses etc. String processing tasks which are performed using regular expressions becomes very complicated because regular expressions language is relatively small and restricted. This mechanism was motivated to avoid silly errors by automated systems, particularly machine learning models. For example, opening salutations, such as "Dear Prabhat", were falsely assigned with instructions by a machine learning model in some runs of cross-validation tests, possibly due to the frequent occurrences of person names in instructions. With this mechanism, "dear __NAME_", a normalized form of "Dear Prabhat", was compared against all such normalized instances in the training data, and false assignment of instructions could be avoided after reviewing emotions assigned to the found training instances, specifically, by confirming more than two-thirds of the found training instances were not assigned with instructions. [20]

Clustering: Clustering is the natural technique used to discover hundreds of feature expressions from text for an opinion mining application. Similarity measures used for clustering are usually based on some form of distributional similarity. There are two main kinds of similarity measures those relying on pre-existing knowledge resources (e.g., thesaurus, and semantic networks) and those relying on distributional properties of words in corpora.

First, a pre-processing pass could build a list of words and phrases that appear frequently in the review of a particular restaurant but are uncommon in the wider corpus. This should find phrases like the name of a dish that many people are talking about. Second, given the narrow domain of the problem, it should also be possible to hand-build a list of common ideas a reader might want to know about, like service, food, and price. Extracting these combined, specific features should lead to purpose-built vectors that form clusters around relevant concepts. [21]

7. Evaluation Measures

All since the problem of grouping feature expressions is a clustering task, two common measures for evaluating clustering are used the study, Entropy and Purity. Below, we briefly describe entropy and purity. Given a data set DS, its gold partition is G = {g_1....,…,g_j....,…g_k}, where k is the given number of clusters. The groups partition DS into k disjoint subsets, DS_1,…, DS_i, …, DS_k.

- Entropy: For each resulting cluster, we can measure its entropy using Equation (a), where $P_i(g_i)$ is the proportion of g_i data points in DS_i. The total entropy of the whole clustering (which considers all clusters) is calculated by Equation (b)
- Purity: Purity measures the extent that a cluster contains only data from one gold-partition. The cluster purity is computed with Equation (c). The total purity of the whole clustering (all clusters) is computed with Equation (d) [22]

$$\text{entropy}(\text{DS}_\text{i}) = \sum_{j=1}^{k} \text{P}_\text{i}(g_i)\log_2 \text{P}_\text{i}(g_i) \quad \text{….Eq.(a)}$$

$$\text{entropy}_\text{total} = \sum_{i=1}^{k} \frac{|\text{DS}_\text{i}|}{|\text{DS}|}\,\text{entropy}(\text{DS}_\text{i}) \text{……Eq.(b)}$$

$$\text{purity}(\text{DS}_\text{i}) = max_j \text{P}_\text{i}(g_i) \quad \text{………………Eq.(c)}$$

$$\text{purity}_\text{total} = \sum_{i=1}^{k} \frac{|\text{DS}_\text{i}|}{|\text{DS}|}\,\text{purity}(\text{DS}_\text{i}) \text{……….Eq.(d)}$$

Similarly for the evaluation of sentiment classification using regular expression is usually measured by precision and recall. Precision is the fraction of relevant retrieved instances, while recall is the fraction of retrieved relevant instances. Therefore precision and recall are based on an understanding and measure of relevance.

PRecall=TP/TP+FN+FP

F-score=2*(precision*recall)/(precision+recall)

Where, TP - number of true positives
TN-number of true negatives
FP - number of false positives
FN - number of false negatives. [20]

8. Tools

A Red Opal is a tool that enables users to find products based on features. The features from customer reviews are used for scoring each product. Opinions on web are analysed and compared using Opinion observer. The product opinions are displayed feature by feature in graph format.

Automation of aggregation sites is done by Review Seer tool. The extracted features are assigned score by Naïve Bayes classifier as positive and negative review. The crawled pages are not classified properly by this tool. Result is displayed in the form of attribute and its score.

Product features are extracted in Web Fountain using beginning definite Base Noun Phrase (bBNP) heuristic. The sentiment lexicon and sentiment pattern database are used to assign sentiments to feature. Sentiment extraction patterns are defined in sentiment pattern database and polarity of terms is defined in sentiment lexicon.

9. Challenges in Opinion Mining

1. Product reviews, comments and feedback could be in different languages (English, Urdu, Arabic, french etc), therefore to tackle each language according to its orientation is a challenging task.
2. As noun words are considered as feature words but Verbs and adjectives can also be used as feature words which are difficult to identify.
3. If a customer-One comments on mobile phone, “the voice quality is excellent” and customer-Two comments, “sound quality of phone is very good”. Both are talking about same feature but with different wording. To group the synonym words is also a challenging task.
6. Orientation of opinion words could be different according to situation. For example “Camera size of mobile phone is small”. Here adjective small used in positive sense but if customer parallel said that “the battery time is also small”. Here small represent negative orientation to battery of phone. To identify the polarity of same adjective words in different situation is also a challenging task.
7. As the customer comment in free format, she can use abbreviation, short words, and roman language in reviews. For example u for you, *cam for camera, pic for picture, f9 for fine,b4, before, gud for good etc.* To deal with such type of language need a lot of work to mine opinion.
8. Different people have different writing styles, same sentence may contain positive as well as negative opinion, so it is difficult to parse sentence as positive or negative in case of sentence level opinion mining .
9. In Bing Liu approach opinion always classified only in two categories positive and negative but Neutral opinion also expressed sometimes. Liu considers only adjective as opinion words but opinion can also expressed as adverb, adjectives and verb. For example “like” is a verb but also an opinion word. His approach finds the implicit features because it extracts the sentences contain at least one feature word. So the features commented by customer indirectly are ignored [17].
10. Lexicon based methods use for opinion mining has not an effective method to deal with context dependent words. For example the word “small” can express the either positive or negative opinion on the product features. For a mobile phone if customer comments that “size of mobile phone is small” this sentence does not show either size is positively opinioned or negatively.
11. To finding of spam and fake reviews, mainly through the identification of duplicates.
12. The comparison of qualitative with summary reviews and the detection of outliers, and the reputation of the reviewer.
13. The combination of opinion with behavior to validate data and provide further analysis into the data ahead of opinion expressed.
14. The continuous need for better usability and user-friendliness of the mining systems.

10. Conclusions

Opinion miming is an emerging field of data mining used to extract the pearl knowledge from huge volume of customer comments, feedback and reviews on any product or topic etc. A lot of work has been conducted to mine opinions in form of document, sentence and feature level sentiment analysis It is examined that now opinion mining trend is moving to the sentimental reviews of twitter data, comments used in Facebook on pictures, videos or Facebook status. In future, Opinion Mining can be carried out on a set of reviews and set of discovered feature expressions extracted from reviews. The state-of-art for current methods, useful for producing better summary based on feature based opinions as positive, negative or neutral is the Expectation Maximization algorithm based on Naïve Bayesian is the most efficient method. The efficiency of EM algorithm can be increased by augmenting it, to reassign classes of the labelled set.

The natural language text can be processed based on machine learning toolkit called as OpenNLP library. The NLP tasks, such as tokenization, part-of-speech (POS) tagging, named entity extraction, parsing, chunking, sentence segmentation, and co reference resolution are provided by Open NLP library. The advanced text processing services are built using these tasks. OpenNLP also includes perceptron and maximum entropy based machine learning.After POS tagging, opinion retrieval can be performed by extracting product candidate feature, related opinion and producing opinion feature pairs. The keywords extracted from Opinion Retrieval Module can be used to perform similarity check with the database dictionary. The similarity check can use semi supervised learning.

References:

1. T. Khushboo "Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm" ,Int. Journal. Computer Technology & Applications, Vol 3 IJCTA | MAY-JUNE 2012.
2. N, Anwer and A, Rashid "Feature Based Opinion Mining of Online Free Format Customer Reviews Using Frequency Distribution and Bayesian Statistics" Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on 16-18 Aug. 2010.
3. B. Seerat and F.Azam "Opinion Mining: Issues and Challenges (A survey)" International Journal of Computer Applications (0975 – 8887) Volume 49– No.9, July 2012.
4. N. M. Shelke, S. Deshpande and V. Thakre "Survey of Techniques for Opinion Mining" International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.
5. D. S. Deshpande " A Survey on Web Data Mining Applications" Emerging Trends in Computer Science and Information Technology -2012(ETCSIT2012) Proceedings published in International Journal of Computer Applications® (IJCA).
6. R. Kohavi and F.Provost "A research on web datamining and its application in electronic commerce
7. ' computational Intelligence and software Engineering 2009. CiSE 2009. International Conference on Date of conference 11-13 Dec 2009.
8. Singh and Vivek Kumar, A clustering and opinion mining approach to socio-political analysis of the blogosphere, Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference.
9. Alexander Pak and Patrick Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining,
10. Zhai Z, Liu B, Xu H, and Jia P, Grouping Product Features Using Semi-supervised Learning with Soft-Constraints, in Proceedings of COLING. 2010
11. Khairullah Khan and Baharum B. Baharudin, Identifying Product Features from Customer Reviews using Lexical Concordance, Research Journal of Applied Sciences Engineering and Technology, 2012
12. Boris Kraychev and Ivan Koychev, Computationally Effective Algorithm for Information Extraction and Online Review Mining, 2010
13. ZhongchaoFei, Jian Liu, and Gengfeng Wu: "Sentiment Classification Using Phrase Patterns", Proceedings of the Fourth International Conference on Computer and Information Technology in 2004.
14. Wilson, T., Wiebe, J. and Hwa, R, "Just how mad are you? Finding strong and weak opinion clauses", Proceeding of National Conference on Artificial Intelligence in 2004.
15. Hiroshi, K., Tetsuya, N., and Hideo, W. "Deeper sentiment analysis using machine translation technology". In Proceedings of the 20th international Conference on Computational Linguistics (Geneva, Switzerland) in 2004.
16. M Fan, G WU "Opinion Summarization of Customer comments" International conference on Applied Physics and Industrial Engineering in 2012.
17. B. Liu "Sentiment Analysis and Opinion Mining", April 22, 2012.
18. Kim, Y. and Myaeng, S., "Opinion Analysis based on Lexical Clues and their Expansion", Proceedings of NII Test Collection for Information Retrieval in 2007.
19. Chuong B Do & Serafim Batzoglou, What is the expectation maximization algorithm?, 2008 Nature Publishing Group http://www.nature.com/naturebiotechnology
20. Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun and Tom Mitchell, Text Classification from Labeled and Unlabeled Documents using EM, Machine Learning, 39, 103–134, 2000.2000 Kluwer Academic Publishers. Printed in The Netherlands.
21. Sunghwan Sohn,1,* Manabu Torii,2,* Dingcheng Li,1 Kavishwar Wagholikar,1 Stephen Wu,1 and Hongfang Liu1, A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes, Biomed Inform Insights. 2012; 5(Suppl. 1): 43–50. Published online 2012 January 30. doi: 10.4137/BII.S8961
22. YuanbinWu, Qi Zhang, Xuanjing Huang, LideWu, Phrase Dependency Parsing for Opinion Mining, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1533–1541, Singapore, 6-7 August 2009. c 2009 ACL and AFNLP