IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.
www.ijiset.com
ISSN 2348 – 7968

An opinion mining tool for mobiles

Raghav S[1], Srushtika Neelakantam[2], RevanaSiddappa Bandi[3]

1, 2, 3 Department of Computer Science and Engineering, Sir M Visvesvarya Institute of Technology, Yelahanka, Bangalore, Karnataka, India

Abstract

Opinion of the others plays an important role in seeking the advice on decision taking. People used to ask the other's opinions orally in earlier days. As internet has been exploded beyond expectations by reaching common men, we are enriched with opinion rich resources like review sites and blogs. With this advent people started getting new opportunities which are associated with some challenges. These challenges include the overloading of information since there are many opinions expressed by different people and manually going through and digesting this huge amount of information lead to some tools which automatically does this. This tool is called sentiment analysis. Sentiment analysis goes through all the opinions about a product in a review site or blog and gives the overall opinion about that product. This tool can be used to take an informed decision for a customer. In this paper we present a tool that finds an overall review of some products.

Keywords: Sentiment analysis, Opinion mining, text extraction, text cleaning.

1. Introduction

Every customer before going to purchase a product seek information regarding that product for the people who bought it and already started using that product. Otherwise he may visit different stores providing the product manufactured by different companies. This is very time consuming for that customer as he has to collect all the features and compare them to take decision of which product he is going to buy. The company which is manufacturing a product may want to know the customer's view on that product. Earlier the customers were given some feed back forms regarding the product and based on those the overall view of that product used to be collected from those feed back forms.

The World Wide Web made these methods of opinion collection obsolete by allowing the people to gather the relevant information for the websites. These days are witnessing the increase in the number of people who are using the internet. There is a rapid growth in the web content also. So many blogs, websites, review sites and discussion forums started emerging. But there is hitch here! The user is overloaded with huge amount of information. Going through each and every review or post in different websites and digesting overall opinion is time consuming, very tedious and daunting task. The user can be provided with a tool which can take a product name and spans over a website or different web sites and extracts overall opinion about that product. The opinion can also be called as sentiment. A sentiment can be defined as the view expressed by a person regarding a product, company or movie etc. A sentiment can be a positive, negative or neutral. Sentiment classification can be considered as a special case of categorization of text problem in which the classification is done according to the attitude expressed by the authors in the blogs or discussion forums etc.

This sentiment analysis can be considered as an extension to the information retrieval since the information which is retrieved using various techniques can be used for sentiment analysis. In recent years sentiment analysis has become a buzz word since people from various fields started using it. Earlier it was used to review movies. Then it was applied in many areas like product reviews, travel reviews, stock market analysis etc. The companies are hiring the people who do this sentiment analysis for tracking the reviews about their products. There are several additional advantages to this approach. First, the people who share their views usually have more pronounced opinions than average, which are additionally influencing others reading them, leading to so called word-of-mouth marketing. Extracting these opinions is thus extra valuable. Second, opinions are extracted in real-time, allowing for quicker response times to market changes and for detailed time-based statistics that make it possible to plot trends over time.

ISSN 2348 – 7968

2. Motivation

The motivation behind this work is, an article in a news paper how the opinions and reviews expressed by the people on some products in the review sites, blogs and discussion forums are affecting the overall sales of their products. An other reason was a malicious attack which was held against ICICI Bank In 2008 the stock value of ICICI was dipping suddenly and many customers were withdrawing their money and started closing their accounts in that bank. Some miscreant people started rumors "Kindly withdraw all your deposits and cash from your account in ICICI Bank as ICICI Bank already rushed to RBI for insolvency.", through the internet It was noticed the rumors are spreading via SMS and online. Getting the SMS data is tricky. There are privacy issues and it cannot get until the cyber crime police is involved. The other option is to capture size-able online postings and use this as proof. With the power of this information, the bank submits it all to the respective authorities. The bank responds to each of the posts.

3. Related work

One of the first attempts in this field was in identifying the *genre* of texts, for instance subjective genres (Karlgen and Cutting, 1994; Finn et al., 2002). The initial approaches to sentiment detection all used linguistic heuristics, explicit list of pre-selected words and other such techniques that require use of experts' knowledge and may not yield the best possible results in all cases as pointed out in Bo Pang et ., 2002. The first attempt to automate the task of sentiment classification was seen in the work of Turney(2002). He used the mutual information between a document phrase and the words "excellent" and "poor" as a metric for classification. The mutual information was determined on the basis of several techniques are used for the opinion mining tasks. To extract opinions, machine learning method and lexical pattern extraction methods are used by many searchers. In 2002, Turney introduced the results of review classification by considering the algebraic sum of the orientation of terms as respective of the orientation of the documents but more sophisticated approaches are introduced by focus on some specific tasks: finding the sentiment of words by Hatzivassiloglou , Wibe , Riloff et al , Whitelaw et al , Dave et al. subjective expression by Wilson et al

Pang and Lee [1] are the first to apply machine learning techniques to text classification problem. During feature selection they have used the Bag-of Words approach and extracted nearly 16000 features. For learning they have used the Naive bayes, Maximum Entropy and Support Vector Machine Algorithm under a 3 Fold cross validation evaluation good observations using bigrams (2 word combinations), POS tagging etc. Lee [1] again extended their previous work in which they extracted only the subjective sentences by filtering the non-subjective ones. Here they extended the data set to 2000 equally distributed reviews and made it standard. They have obtained comparable performances over the previous one. Konig and Brill used a hybrid classifier which works in two steps; in the first phase they used a pattern based classifier and if the document is not classifiable at first phase, it is sent to general learning based classifier at second step . In India many companies like Valuepitch Interactive, Pinstorm are doing sentiment analysis and they have many clients across India who want to keep track of the sentiments about their products,

4. Proposed System

We used the review of mobile phone based on its features like overall phone, sound, camera, screen, battery from the website www.gsmarena.com. This is done in two ways online data analysis and offline stored data analysis. First based on the input (in this case it is phone name) we had retrieved web pages. Since the pages contain irrelevant data, text is cleaned .Based on bag of good words and bad words the sentiment is extracted. In this we get overall reviews, specification wise sentiment and overall product sentiment.

ISSN 2348 – 7968

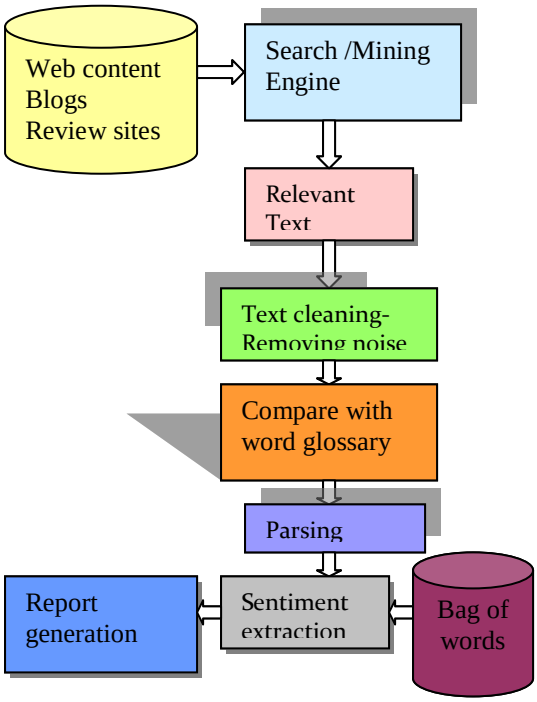

Fig 1: Basic Architecture

The steps are explained below:

1. If the application is integrated into a general-purpose search engine, then one would need to determine whether the user is in fact looking for subjective material. This may or may not be a difficult problem in and of itself: perhaps queries of this type will tend to contain indicator terms like “review,” “reviews,” or “opinions,” or perhaps the application would provide a “checkbox” to the user so that he or she could indicate directly that reviews are what is desired; but in general, query classification is a difficult problem — indeed, it was the subject of the 2005 KDD Cup challenge. We gathered data from internet that solely based on the (SOR) Subject of Reference (e.g. nokia). We used web mining techniques to gather all web pages where the SOR is mentioned.

ISSN 2348 – 7968

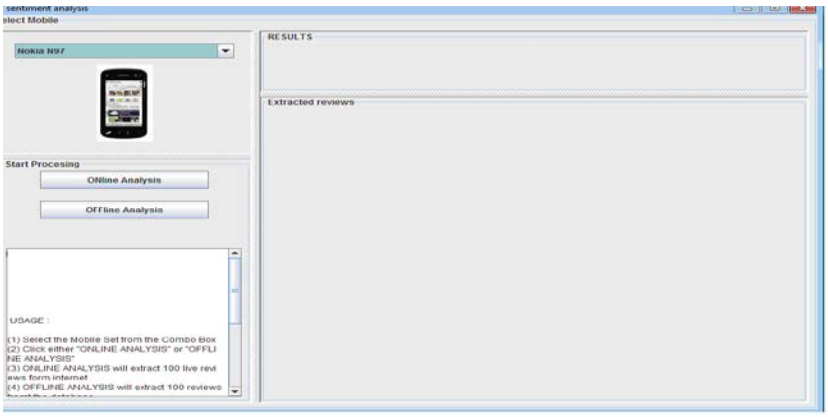

Fig2: The user can select the mobile and offline/online analysis

In the above picture SOR is Nokia phone.

2. Besides the still-open problem of determining which documents are topically relevant to an opinion-oriented query, an additional challenge we face in our setting is simultaneously or subsequently determining which documents or portions of documents contain review-like or opinionated material.

Text Extraction can be done in several data mining or text mining techniques starting from simple 'keyword matching' to 'DOM structure mining' to 'neural networks' methods. The major challenge here is that web documents are highly unstructured and no single method can give 100% clean text extraction for all documents. We have used pattern matching by searching for some characters

3. Text Cleaning is mostly heuristic based and case specific. By this we mean is to identify the unwanted portions in the extracted contents from Step 2 with respect to different kinds of web documents (e.g. News article, Blogs, Review, Micro Blogs etc) and then write simple cleanup codes based on that learning, which will remove such unwanted portions with high accuracy.

4. We then check for the presence of specifications. In our case specifications are overall sound quality, screen quality, camera quality etc.

5. Then we had searched for the words that contain words which lead the sentence to bad, good or neutral. For every specification, we checked for words which are present either of the two sets of good and bad words. We classified words into two classes (positive or negative) and counted on overall positive/negative score for the text. If the documents contain more positive than negative terms, it is assumed as positive otherwise it is negative. These classifications are based on sentence level classification.

6. The results are shown in pie chart and detailed reviews. The results are shown for individual specifications as well as for overall sentiment for the product.

ISSN 2348 – 7968

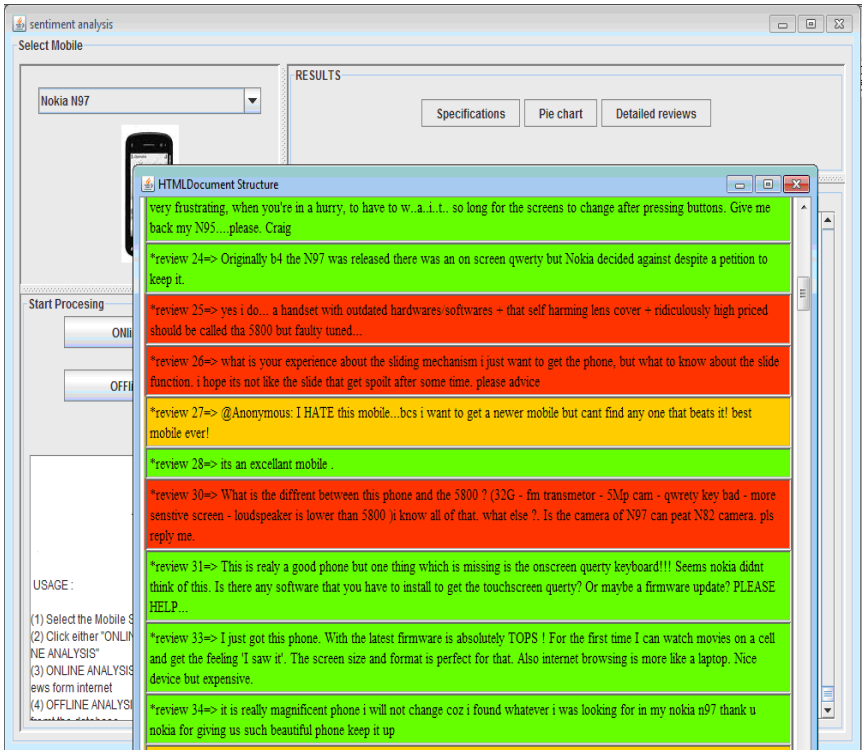

Fig 3: Detailed reviews

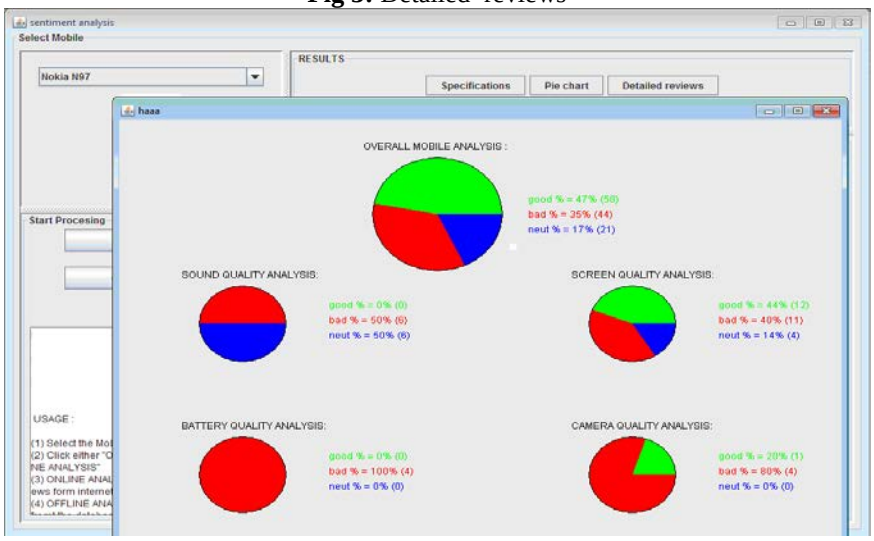

Fig 4: Piechart showing results

5. Performance measure

To evaluate sentiment classification system, we use the precision and recall measures in the following ways.
Precision = number of correct positive predictions/ number of positive predictions
Recall = number of correct positive predictions/number of positive examples

ISSN 2348 – 7968

In the context of classification tasks, the terms true positives, true negatives, false positives and false negatives are used to compare the given classification of an item with the desired correct classification. This is illustrated by the table below:

	correct result / classification	
	E1	E2
Obtained result / classification	Tp (true positive)	fp (false positive)
	Fn (false negative)	tn (true negative)

Sensitivity= 92 and specificity=72, Precision=88.4, recall=92

6. Limitations

In the Indian web space, consumer comments are growing in blogs, forums, review sites and twitter. Indian consumers tend to express their emotions in local languages like Hindi etc. They use a Hindi word in English font - 'ICICI Chor hain' for instance. A search in Google for the same phrase will throw up more than 6000 links. The other common languages used are Telugu, Tamil, Bengali and Kannada.

Sentiment analysis will be inadequate if these expressions are not captured and analyzed. The problem is these words are not part of any standard dictionary and hence identifying the existence itself poses a challenge. Another area is the creative freedom with which people can express these sentiments in different spellings. Together as the web usage grows (internet penetration in India is less than 4% currently and growing at a rate of 25% +), the problem gets bigger.

7. Conclusion And Future work

We have done the work and created and tested. By using this approach we can view the strength or weakness of the products or objects more detail and we hope will be useful for further development and improvement of the products or objects. Further development of this approach is still ongoing since our work deals with only mobiles and we have classified sentiment based on set of words that pertain good or bad .We Also the system is not 100% accurate as no system (especially sentiment system) is.

References:

[1] Opinion mining and sentiment analysis
Bo Pang1 and Lillian Lee2 Vol. 2, No 1-2 (2008) 1–135
[2] Bo Pang and Lillian Lee. Using very simple statistics for review search: An exploration. In Proceedings of the International Conference on Computational Linguistics (COLING), 2008. Poster paper.
[3] Wiebe J.M., “Learning subjective adjective from corpora”,AAAI-2000, 2000.
[4] Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan.
Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing.
[5] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003.
[6] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL), pages417–424, 2002.
[7]Ch.VenkataRamana, CEO,Valuepitch Interactive ,Mumbai
[8] Bo Pang and Lillian Lee, A Sentimental education:Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, Proceedings of ACL, 2004.

ISSN 2348 – 7968

[9] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, Thumbs Up? Sentiment Classification Using Machine Learning Techniques, Proceedings of EMNLP 2002,pp 79-86.
[10] Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke. Using WordNet to measure semantic orientation of adjectives. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), volume IV, pages 1115-1118. European Language Resources Association, Paris, 2004.

[11] Osgood, C. E., G. J. Succi, and P. H. Tannenbaum, 1957. The Measurement of Meaning. University of Illinois Press, Urbana IL.
[12] George Forman An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research 2003, pages 1289-1305.
[13]Y. Li, Z. Zheng, and H. Dai, "KDD CUP-2005 report: Facing a great challenge," *SIGKDD Explorations*, vol. 7, pp. 91–99, 2005.
[14]K. Moilanen, and S. Pulman. The good, the bad, and the unknown: Morphosyllabic sentiment tagging of unseen words. Proceedings of ACL-08:HLT, pp. 109–112, 2008.
[15]S.-M. Kim, and E. Hovy. Determining the sentiment of opinions. Proceedings of Conference on Computational Linguistics, pp. 1367–1373, 2004.
[16] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Textual affect sensing for sociable and expressive online communication. Proceedings of 2nd International Conference on Affective Computing and Intelligent Interaction, pp. 220–231, 2007.
[17] Chenghua Lin, Yulan He, Richard Everson, and Stefan R¨uger,"Weakly-supervised Joint Sentiment-Topic Detection from Text",To be published by IEEE Transactions on Knowledge and Data Engineering-2011
[18] Zhongwu Zhai, Bing Liu, Jingyuan Wang, Hua Xu and Peifa Jia,"Product Feature Grouping for Opinion Mining Using Soft-Constraints and EM", To be published in IEEE Intelligent Systems-2011