Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 73 (2014) 26337-26341

ISSN: 2229-712X

Opinion mining of products using web

Unnamalai.K

Department of Computer Science, Dravidian University, Kuppam, Andra Pradesh, India.

ARTICLE INFO

Article history:
Received: 21 June 2014;
Received in revised form:
25 July 2014;
Accepted: 9 August 2014;

Keywords
Artificial intelligence,
Natural language processing,
Opinion mining.

ABSTRACT

The evolution of web 2.0 has made social media networks more prevalent. Forums, blogs, tweets and posts, have become an important media for internet users to share views. With the evolution of Natural Language Processing (NLP), sentiment analysis is being essentially used as a means to determine the attitude of a person with respect to some topic or the overall contextual polarity from these inputs. The attitude may be his or her judgment or evaluation, effective state mind or the intended emotional communication. As e-commerce is becoming increasingly popular, the number of customer reviews that a product receives grows rapidly, which is being encouraged by the merchants selling the products. For a popular product, the number of reviews can be in hundreds or even thousands. This paper describes the concepts of sentiment analysis from unstructured text, looking at why they are useful and what tools and techniques are available. It also focuses specifically on a novel feature based opinion extraction scheme demonstrated with key open-source tools and applications and the problems associated with opinion detection in social media have also been analyzed.

Introduction

Opinion mining is a language processing which helps us to track the mood of the public about a particular product. It is also known as sentiment analysis, which involves collection , processing and development of opinions about the product that is the subject matter of analysis from blog posts, tweets, reviews etc. Machine learning is the basic process involved in opinion mining which is basic Artificial Intelligence(AI).

Opinion mining tends to be useful in several ways. For instance, if you are in business, the mining process aides in judging the probability of success of a new product, to be launched. In addition, this also helps to determine which versions of a product or service are popular and even identify which demographics, liking or disliking particular features. For instance, a review might be highly positive about the over all product, for example, a mobile phone, but be specifically negative about the details of the in-depth specification.

There are several challenges in opinion mining. The first being, a word that is considered to be positive in one situation may be considered negative in another situation. For instance, if the word "high" is taken, in two different contexts, if a customer says a laser jet printer's print quality and print speed are "high", that turns out to be a positive opinion. But if another customer says the same printer's cost is "high" that turns out to be a negative opinion. The context of the opinion sought to be mined will vary and the same cannot have an universal yardstick. The opinion mined will have to be applied in line with the context and cannot be read in isolation so as to convey the relevant meaning as stated above.

Sentiment Analysis [15] is used to understand the polarity of the given text in a document, sentence, or feature/aspect level. The polarity can be based positive, negative, or neutral. Further the sentiment can be culled out from texts, mono syllables etc. For example emotional states such as "angry," "sad," and "happy." Sentiment analysis cannot be constrained only to written text alone but will also have to take into considerations such as emotion icons and words like "Like", "Dislike", "No Comments" etc., which need not opinions coined by the person passing the opinion, but the author of the text may provide such option which may be or may not be subscribed by the persons passing such comments. It should also be understood that whenever any opinion is passed by person about a feature of any product, it is only about the feature and not the product itself. Hence it is necessary to assimilate and understand opinion about the feature in the product only. However natural language processing [24] can also be used understand the exact opinion of the person passing such opinion. However it is not without its own difficulties, which compels the process of opinion mining, for example the terms "y", "yes", "ya" and the likes mean YES and "n" , "no" ,"ne" etc. means NO and the same will have to be culled out of the entire spectrum of opinions.

Turney [19] and Pang [8] have implemented different methods for detecting the polarity of product reviews and movie reviews respectively. The work that has been carried out is at a theoretical level because it is at the document level. Pang [10] and Snyder [5] attempted to classify a document's polarity on a multi-way scale. Pang[10] expanded the basic task of classifying a movie review as either positive or negative to predicting star ratings on either a 3 or a 4 star scale. Snyder [5] performed a detailed analysis of restaurant reviews. These review were used to predict ratings for various aspects like food, ambiance etc. of the given restaurant.

Thelwall et la [22] used another method for determining opinion is to use scaling system where opinionated words are given a score on -5 to +5 scale (negative to positive) and a piece of unstructured text is analyzed using natural language processing, determines how it relates to the concept and then given a score on the scale. Alternatively, texts can be given a positive and negative sentiment strength score if the goal is to determine the sentiment in a text rather than the overall polarity and strength of the text.

At times the process of sentiment analysis will have to be based on subjectivity identification / objectivity identification. The task is to classify the given sentence [3] into one of two

Tele:
E-mail addresses: unnamalaik@gmail.com

classes: objective or subjective. Identifying this is sometimes can be more difficult than polarity classification [21]: the subjectivity of words and phrases may depend on their context and an objective document may contain subjective sentences (e.g., a news article quoting people's opinions). Moreover, as mentioned by Su, [13] results are largely dependent on the definition of subjectivity used when annotating texts. However, Pang [9] showed that removing objective sentences from a document before classifying its polarity helped improve performance.

Andrea Esuli [4] attempted to predict the orientation of subjective adjectives by analyzing pairs of adjectives. They can be conjoined by and, or, but, either-or, or neither-nor. They are extracted from a large unlabeled document set. The idea is that the act of conjoining adjectives is subject to linguistic constraints on the orientation of the adjectives involved; for example and usually conjoins adjectives of equal orientation, while but conjoins adjectives of opposite orientation.

Yun-Qing Xia et al in 2007[26] proposed a novel opinion mining approach which uses a Unified Collocation Framework (UCF) which incorporates attribute-sentiment collocations and their syntactical features. In UCF POS tags are used for both attribute and sentiment keywords. UCF is defined as (UCF) {[attribute], <pos_attribute>, [sentiment_keyword], <pos_sentiment_keyword>, <polarity>}.The learning component of the UCF uses the OPINMINE corpus to find all the general collocations. In the working component the review text is split into words and POS tag is fixed to the words and OOV opinions are found. Performance is evaluated for all the four approaches. The SVM classifier is used in the UCF method. The authors [26] also generates a graph where terms are nodes connected by "equal-orientation" or "opposite-orientation" edges, depending on the conjunctions extracted from the document set. A clustering algorithm then partitions the graph into a Positive cluster and a Negative cluster, based on a relation of similarity induced by the edges.

Turney and Littman[23] determine term orientation by bootstrapping from two small sets of subjective "seed" terms (with the seed set for Positive containing terms such as good and nice, and the seed set for Negative containing terms such as bad and nasty). Their method is based on computing the point wise mutual information (PMI) of the target term "t" with each seed term "ti" as a measure of their semantic association. For a given a target term t, its orientation value is O(t). The value tends to be positive means positive orientation, and higher absolute value means stronger orientation. It is given by the sum of the weights of its semantic association with the seed positive terms minus the sum of the weights of its semantic association with the seed negative terms. For computing PMI, term frequencies and co-occurrence frequencies are measured by querying a document set by means of the AltaVista search engine with a "t" query, a "ti" query, and a "t NEAR it" query, and using the number of matching documents returned by the search engine as estimates of the probabilities needed for the computation of PMI. Further the author extended the above approach to classify Movie reviews and Automobile reviews with few changes to the existing approach. This method classifies the review as recommended or not recommended. Initially he identifies the adjectives and adverbs in the review and uses a part-of-speech tagger to mark the phrase. The adjectives, adverbs, nouns are pre-ordered and accordingly the tags are associated with them.

Proposed Methodology

Sentiment analysis is imbibed in social media, whether the same is understood in the strict sense or otherwise. Any opinion in the internet will be construed as an opinion in the social media. Today social media mainly in categorized into blogs, sites, forums, discussion boards, twitter, Facebook and the likes. It is but a very simple exercise undertaken by most people before buying any product is to type the name of the product, place etc. and conduct a search on the internet and then browse the opinion of various people to make up the mind as to the course action that is to be taken. People eventually make decisions based on the various opinions, ads, reviews, ratings etc. realizing that social media is the real place where opinions are given without any fear or favour, industry has increasingly started looking forward to it for improvement of its products and services. This has necessitated in eliminating the unnecessary comments and looking into the actual communication that is sought to be opined. Many are now looking to the field of sentiment analysis. [25]

One step towards this aim is accomplished in research. Several research teams in universities around the world currently focus on understanding the dynamics of sentiment in e-communities through sentiment analysis [3] The Cyber Emotions project, for instance, recently explored about the impact of negative emotions in redirecting and driving a discussion on chosen topics on the social network.[20] The study of Sentiment analysis henceforth, aided in arriving at a conclusion as to why certain e-communities cannot survive in the virtual world and eventually fade away, for instance MySpace, whereas a few others take the social media by storm and tend to flourish beyond imaginary lines, like Facebook.

Social media provides a wealth of information about a user's behavior and interests. The sentiment of the user can be of three types, they are explicit(eg: John likes tennis, swimming and classical music), implicit (people who like skydiving tend to be big risk-takers)and associative(people who buy Nike products also tend to buy Apple products).The target text can be Users, documents(blogs), Sentences (paragraphs, chunks of text?), Predetermined descriptive phrases (<ADJ N>, <N N>, <ADV ADJ>, etc.), Words and Tweets/updates.

The first step in this paper is identifying features of the product that customers have expressed their opinion. A product feature attributable to the product that has been commented on by the customer. The feature can be explicit or implicit. For example, "battery life" in the following two opinion Sentences is an explicit feature as it appears explicitly:

- "The battery life of this phone is too short"
- "Battery life too short"

In the next two example opinion sentences "Size" is an implicit feature size is as it does not appear in each sentence but it is implied:

- "This phone is too big to carry"
- "Too big"

The second step is to determine whether the opinions for every feature are positive, negative or neutral. The opinion holder can expresses a positive or negative opinion for each feature. One sentence may be used to express opinions of more than one feature. It is not necessary that all the opinion has to be positive or negative. The sentence can express both positive and negative, as the following two sentences shows:

- "The software is good, but the battery life is short"
- "Good Bluetooth connectivity, and long battery life"

The next step is to Group the synonyms of features (as different opinion holders may use different words or phrase to express the same feature). For example, one reviewer may use "photo", but another may use "picture". Synonym of features should be grouped together. The last step is Opinion Summarization. One simple way is to produce a feature-based summary of opinions on the object [16].

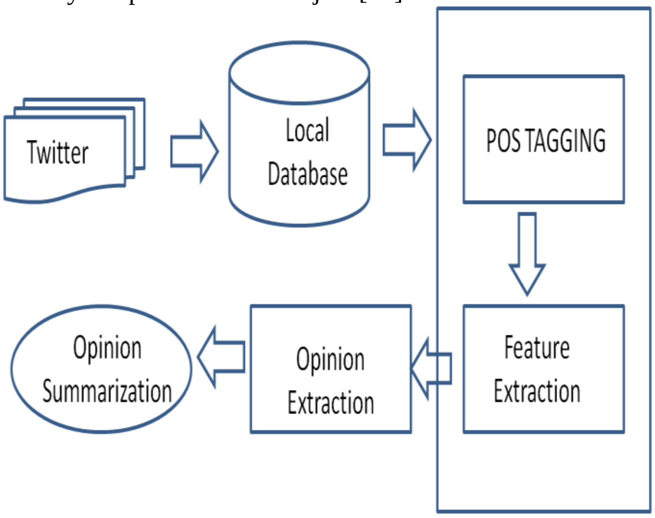

Fig 1. System architecture

Implementation

Twitter is considered as the source. The tweets about a particular product can be downloaded to a local database by using Twitter API 4J. Tweets on a particular timeline can be filtered.

POS tagging

Once the tweets are selected the next step is to POS tagging. POS tagging is important as this process identifies each part of sentence that needs to be processed

The following are the various methods of tagging. They are Rule-Based POS tagging, Transformation-based tagging, Stochastic (Probabilistic) tagging .The main focus lies in the processing of adjectives. The tweets are tagged using a Stanford POS tagger here. We run each review through a part-of-speech Stanford (POS) tagger. This appends the respective POS to each word, and then appends the dictionary form of the word, known as its "lemma." For example, "Samsung Galaxy's battery life is about 9hrs. This phone has got good features for lesser price. It also has very good camera performance" becomes "Samsung/NNP Galaxy/NNP's/POS battery/NN life/NN is/VBZ about/IN 9hrs/CD./. This/DT phone/NN has/VBZ got/VBN good/JJ features/NNS for/IN lesser/JJR price/NN./."

Once the texts are tagged, parser is used to generate the parse tree. The tree having three sub nodes i.e., "NP"," VP" and "."Each sub nodes has a sub tree. NP sub tree contains nodes of nouns and adjective whereas the VP sub tree contains the nodes of verbs and adverbs.

Feature extraction

Once POS Tagging is completed, features of the products from the tweets are to be extracted. Classifying evaluative texts at the sentence level does not tell what the opinion holder likes or dislikes. A positive sentence on a product does not mean that the opinion holder has positive opinions on all aspects or features of the product. Likewise, a negative comment does not mean that the opinion holder dislikes everything about the product. In a review, the opinion holder typically writes both positive and negative aspects of the feature, even though the general sentiment on the product may be positive or negative. To identify the detailed aspects, feature extraction becomes necessary. So Identifying object features is the next step. For instance, in the sentence "The camera quality of this phone is amazing," the object feature is "camera quality". This can be implemented [7] supervised pattern method or [16, 14], an unsupervised method is used method can be proposed. This paper basically extracts frequent nouns and noun phrases as features, which are generally genuine features of products. Clearly, much information extraction techniques are also applicable, e.g., conditional random fields (CRF), hidden Markov models (HMM), and many others.

Opinion Extraction

The main goal of this paper is to find the sentiment of the user. For each review extract the opinion of the features. Opinion Extraction can be done at different levels. First and foremost as level 1 we can aggregate all positive opinion and negative opinion of all features of the product to show an overall customer opinion on the product. As a next level, we can focus on each main feature or component of the product, e.g., "battery", "memory" and "touch screen", and generate its Positive and Negative or neutral opinion, so that we can have number of positive or negative opinions on each feature . As a last level, we can study specific problems of each feature, e.g., "the mobile start-up time is higher" and "the mobile battery's standby time is very less". This paper basically aims to work at level 1 and level 2, which are often sufficient. Details at level 3 and beyond are too specific and are studied by human analysts. The word with JJ tag (ADJECTIVE) has to take and the weight of adjective from Sentiwordnet Database. It is the polarity of the tweet. The word "data" is plural, not singular

Grouping of synonyms feature

As the same object features can be expressed with different words or phrases, this module groups those synonyms features together. We measure similarity between two terms by looking at the similarity among the individual words in the terms. In what follows, we introduce three metrics for word similarity:

The process of stemming refers mainly to the reduction of inflected words to their respective stem. This is in general, the base or root form, more prominently a written word form. The stem can be unique and not necessarily identical to the morphological root of the chosen word. The usual convention is that the related words need to necessarily map to the same stem. The stem may not happen to be a valid root by itself. The [14] Crude Factors (Vi) are stemmed and compared with [14] User Defined Factors (Wj), if a match is found it is taken.

$$\text{strmatch}(Vi\ ,\ Wj) = \begin{cases} 1 \text{ if } Vi \text{ matches } Wj \\ 0 \text{ otherwise} \end{cases}$$

a synset or synonym set is defined as a set of one or more synonyms that are interchangeable in some context without changing the true value of the proposition in which they are embedded[14].Find the synset of the CF using WORDNET 3.0 , if a match is found with UDF it is taken or if it is a Polysemous words which belong to more than one synset[22] synscore is calculated as

$$\text{synscore}(Vi,Wj) = \begin{cases} 1 \text{ if syns}(Vi)\ \text{syns}(Wj) \\ 0 \text{ otherwise} \end{cases}$$

Similarity Score [13] sm can also be calculated. This method metric also uses WordNet and requires both POS and sense information. The similarity measures are implemented as PERL module Word-Net:: Similarity [18].

simscoresm(Vi ,Wj) = sim(Vi,Wj) / max(sm)

Orientation summarization

The steps involved in obtaining individual opinion have been integrated to achieve the overall opinion for a product. The above stated process has to be repeated for all tweets and polarity of each tweet is counted as positive, negative and neutral for each feature respectively. Features extracted can be frequent and infrequent. Basically frequent features are considered in this paper and the results can be summarized as overall and with features.

Result

Precision Recall is calculated for the extracted tweets and the table is listed below. The system that has been developed for mining public opinion for product oriented tweets is implemented successfully. This system is found to provide good performance for varying kinds of tweet.

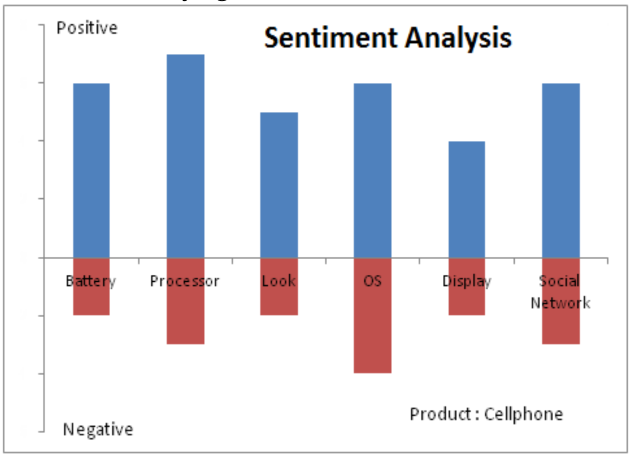

Fig 2 Opinion for features on a particular product

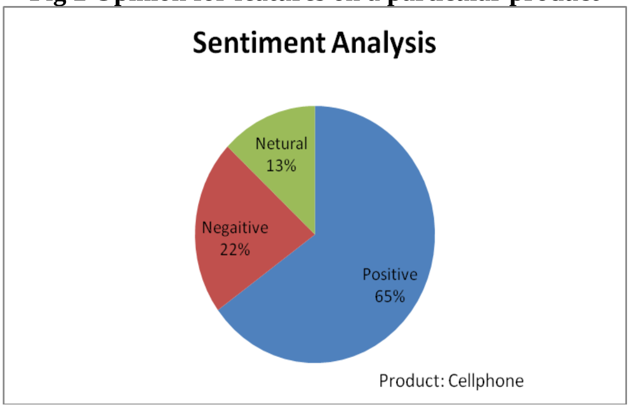

Fig 3. Opinion on a particular Product

Precision table

Features	Recall(%)	Precision(%)	Error-rate(%)
Battery	65.67	74.58	22.39
processor	66.26	66.67	33.13
Look	64.86	72.37	24.76
OS	64.23	72.28	25.06
Display	68.72	78.34	18.99
Social Network	60.96	74.79	20.55
Avg	65.12	73.17	24.15

Conclusion

This work can be further extended to extract infrequent features of the product specification and mine opinion from those features.

References

Alec Go, A Go, L Huang, R Bhayani, "Twitter Sentiment Analysis", Volume: 2009, Issue: June, Publisher: Association for Computational Linguistics, Pages: 1-17.

Andrea Esuli, "Opinion Mining" Language and Intelligence Reading Group.

Andrea Esuli, Fabrizio Sebastiani, 2005, "Determining the Semantic Orientation of Terms through Gloss Classification", Conference on Information and Knowledge Management, 2005.

Andrea Esuli, "Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications",2008

Benjamin Snyder; Regina Barzilay (2007). "Multiple Aspect Ranking using the Good Grief Algorithm". Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL). pp. 300–307.

Bing Liu (2010). "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

Bing Liu; Minqing Hu and Junsheng Cheng (2005). "Opinion Observer: Analyzing and Comparing Opinions on the Web". Proceedings of WWW 2005.

Bo Pang; Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". Proceedings of the Association for Computational Linguistics (ACL). pp. 271–278.

Bo Pang; Lillian Lee and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.

Bo Pang; Lillian Lee (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". Proceedings of the Association for Computational Linguistics (ACL). pp. 115–124.

Carenini, G., Ng, R. and Zwart, E. Extracting Knowledge from Evaluative Text. Proceedings of the Third International Conference on Knowledge Capture (K-CAP'05), 2005.

Condliffe, Jamie. "Flaming drives online social networks ", NewScientist, 2010-12-07. Retrieved on 2010-12-13.

Fangzhong Su; Katja Markert (2008). "From Words to Senses: a Case Study in Subjectivity Recognition". Proceedings of Coling 2008, Manchester, UK.

G.A Miller "WordNet: A Lexcial Database for English"

Michelle de Haaff (2010), Sentiment Analysis, Hard But Worth It!, CustomerThink, retrieved 2010-03-12.

Minqing Hu; Bing Liu (2004). "Mining and Summarizing Customer Reviews". Proceedings of KDD 2004..

Pang, Bo; Lee, Lillian (2008). "4.1.2 Subjectivity Detection and Opinion Identification". Opinion Mining and Sentiment Analysis. Now Publishers Inc.

Patwardhan and T. Pedersen. Cpan module wordnet::similarity, 2003.

Peter Turney (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics. pp. 417–424. arXiv:cs.LG/0212032.

Popescu, A.-M. and Etzioni, O. Extracting Product Features and Opinions from Reviews. Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP'05), 2005.

Rada Mihalcea; Carmen Banea and Janyce Wiebe (2007). "Learning Multilingual Subjective Language via Cross-Lingual Projections". Proceedings of the Association for Computational Linguistics (ACL). pp. 976–983.

Thelwall, Mike; Buckley, Kevan; Paltoglou, Georgios; Cai, Di; Kappas, Arvid (2010). "Sentiment strength detection in short informal text". Journal of the American Society for Information Science and Technology.

Turney Littman, “Unsupervised Learning of Semantic Orientation Unsupervised Learning of Semantic Orientation” NRC Publications Record / Notice d'Archives des publications de CNRC, 2002

Unnamalai K, “Sentiment analysis of Products using Web”, Proceedding of ICMOC, 2012, Procedia Engineering, volume 38, pages 2257-2262.

Wright, Alex. "Mining the Web for Feelings, Not Facts", New York Times, 2009-08-23. Retrieved on 2009-10-01.

Yun-Qing Xia1, Rui-Feng Xu2, Kam-Fai Wong3, Fang Zheng,” The Unified Collocation Framework for Opinion Mining”.