Sentiment Classification for Dynamic Opinions using Natural Language Processing

Jayanag B[1]

Sr. Asst Professor, Department of CSE,

V.R.Siddhartha Engineering College.

Dr. K.V. Sambasiva Rao[2]

Principal,

M.V.R College of Engineering, Paritala.

Abstract

The opinions and reviews posted in the websites are analyzed mostly based on static data. Till date, the methods applied to determine whether the opinions are positive or negative. But, now a days, as products are increasing day by day and a huge amount of data reviews are available online. So, this mechanism not considers the opinions at comprehensive level.

This paper proposes a new method for classifying sentiments that are extracted dynamically from websites. Reviews of various products and comments from social networking sites are aggregated using Bayesian Networks and Natural Language processing Techniques. An improved stemming algorithm is also proposed in this paper. The proposed model input a group of comments and formulates ranks to the reviews for the products. Thus, user can easily estimate the quality of the product based on the comments posted in social networks.

1. "Introduction"

Sentiment classification is a technique to classify people's opinions in product reviews, blogs or social networks. It has different usages and has received much attention from researchers and practitioners. The availability of public opinion over the Internet and in face to face conversations, coupled with the need to understand and mine these for end applications, has motivated a great amount of research held in recent times. Researchers have explored a wide array of knowledge resources for opinion analysis, from words and phrases to syntactic dependencies and semantic relations. Sentiments are also known as opinions/comments/messages.

The present opinion mining is done statically only for a small set of data and the dependencies in the opinions are not considered for summarization. While many methods have recently been introduced, they fall short of evaluating the quality of the results they produce in a systematic way, which is mostly caused by the lack of publicly available test collections. Ratings or voting's are taken only for a period of time for that site only. User needs to wait till these ratings or voting's get closed and then will come to a conclusion of which product is best. But if the user wants to know the results in another site it might be an overhead to the user and these results might vary from site to site. And also any of the available opinion mining systems doesn't consider the dependencies associated in the opinions. A system that could automatically process the comments and generate a generalized result out of the list of comments posted about a product by considering the dependencies could be useful to give a brief synopsis of the product.

In this study, we are interested in product feature based sentiment analysis. In other words, we are more interested in identifying the opinion polarities (positive, neutral or negative) expressed on product features than in identifying the opinion polarities of reviews or sentences considering the dependencies in the opinions. Several studies have applied unsupervised learning to calculate sentiment scores of product features. Although many studies used supervised learning in document-level or sentence-level sentiment analysis, we did not come across any study that employed dealing with the dependencies in the opinions and product features for sentiment analysis. To find the sentiments of the opinions we used natural language processing techniques and to resolve dependencies we considered Bayesian network. The proposed architecture is separated in to two phases

1) Preprocessing: The data is extracted and filtered using web crawls tools and natural language processing techniques.
2) Ranking: The features are identified using POS tagging and the overall sentiments and dependencies are predicted using Bayesian network.

2. "Related work"

Researchers in sentiment analysis have focused mainly on two problems– detecting whether the text is subjective or objective, and determining whether the subjective text is positive or negative. The techniques relied on two main approaches: unsupervised sentiment orientation calculation, and supervised and unsupervised classifications.

Abbasi et al. [1] compared various feature representations for analysis and proposed a Support Vector Regression Correlation Ensemble (SVRCE) method for improved classification of intensities. They combined all together with affect correlation information to enable better prediction of emotive intensities. There considerations are web, blogs and articles. They considered n-gram based features and word is used for awarding the scores to the opinions. But they have not considered the inter dependencies between the words.

Minqing Hu et al. [3], work is closely related to our work. Their task is performed in 3 steps: Mining product features, identifying opinion sentences and identifying the positive or negative ness of each sentence, summarizing the results. Their system doesn't have a pre-processing phase, the review data base is directly sent to POS Tagging phase for feature identifications where the unnecessary words will also be considered and the time for entire processing will be increased. So including the pre-processing phase the data set size can be reduced so that the system accuracy can be increased. And also in predicting the opinion orientations they used a Sentence Orientation procedure but using the WordNet based score and Bayesian networks the results could be better.

Michael Gamon et.al. [5], proposed a clustered model called pulse. It works on both sentiment detection and topic detection free form customer opinions .They worked for both sentiment detection and topic detection performed at the sentence level. They used car reviews that are related to different models. In order to train pulse model, a bootstrap classifier is used. This takes little set of seed words from the reviews, which are annotated manually with their sentiments either positive or negative.

Dave.K et al. [11], worked on sentiment classification (either positive or negative) of product reviews by training the classifier with the corpus of amazon binary rated (thumbs up and thumbs down) and scalar rated(1-5) comments. They segmented each document in to sentences with mini pars linguistic parser to classify the opinions and also tagged each token with parts of speech.

Trivedi K et al. [6], Classification of sentiments using naïve Bayesian algorithm has been proposed. They sustain tri,bi and unigram words of data labeled with polarity but this process is static method which increases the effort to analyze the overall product or a brand reviews. When the words are matched to the database and order of them does not change. So there should be closeness over the order.

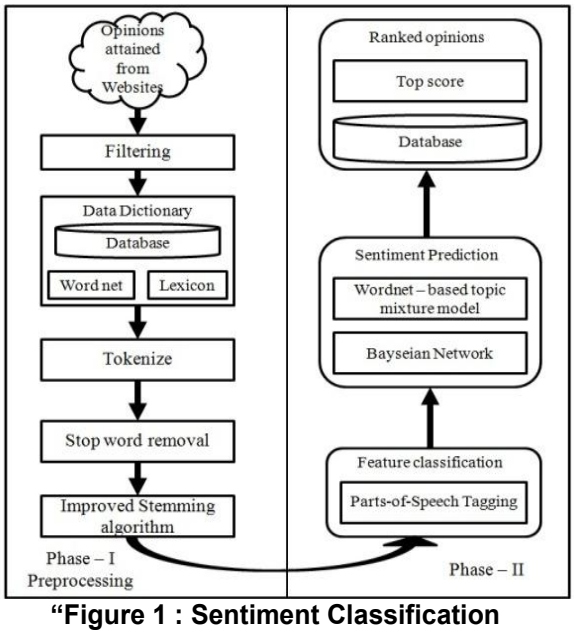
Akshat B et al. [7], various approaches of NLP like N-gram and POS tagged n-gram matching for identification of sentiments. Here all the combinations are used along with some parts of speech to boost the performance of the classifier. They developed an algorithm and new scoring function to classify reviews of both product and movie data sets.

3. "Proposed system"

Now a days, online shopping is very familiar and most of the users are interested to buy products online and most products will have similar features. So, a user cannot decide which type of product suits to his requirements. The reviews given by different users in social networking site would help the customer decide the quality of the product and its features ranking.

Here, we propose an architecture where a tool is used to crawl the data dynamically from various websites. The specific data relevant to a particular product is reviewed and filtered. These sentences are stored in a massive database and are pre-processed. These opinions database may alter time to time as comments are posted through online by users from various parts of the world. Textual information available from web is classified as facts and opinions. Facts are objective statements on news, people, etc while opinions are subjective sentences consisting of attitudes, emotions and feelings. Opinions are generally available online in various domains. But, product based comments along with movie related content are most commonly used to mine the knowledge. An improved architecture is proposed and a brief proposed architecture can be found [9].

It is necessary to understand basic concepts before coming to methodology and working of sentiment classifier.



"Figure 1 : Sentiment Classification Architecture"

A) Preprocessing

In this phase, the data is crawled from web by downloading documents and following links from various pages and the data is filtered. The filtered comments are stored in to an Excel sheet, which are stored dynamically and the comments databases are refreshed in a frequent interval of time. Two main functions of a crawler are fetching static html pages of a given URL and parse the html tags in each page to extract the desired content of the page. Many tools have been developed to crawl contents of web. Some of them are CNET web scrapper, AMAZON reviews downloader, parser for Amazon reviews, web text extractor software etc that are available online.

English dictionaries are repositories of millions of words and their meanings. WordNet [10] is an on-line lexical reference system. WordNet shows noun, adjective, adverb and verb synonym sets for each input word. WordNet is not only a dictionary of words but it also takes care of word sense disambiguation as each English word sense or a part of speech differs based on the context and meaning. Words can be of any sense which is dependent on the meaning that the word gives to a sentence. So, Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms. In lexicon dataset we have constructed our own data set for stop word removal. Some of them like 'a', 'all', 'is', 'on', 'when', 'the', 'those', 'they', 'I', 'we', 'which', 'in'.

Next step is tokenization. Here, words given are identified by using natural language toolkit tokenizers like the below said example.

The laptop is good.

|The| |laptop| |is| |good| |.|

Further, the insignificant words are rejected. This is done to reduce the size of the opinion by using stop word removal technique. Finally, the opinions are stemmed by using improved stemming algorithm enhances the stemming process by identifying the errors in porter's stemming algorithm [12]. The inaccuracies encountered during the stemming process also overcome corresponding solutions are proposed.

B) Feature Identification

To classify a document, we generally start with a very large number of words that need to be considered, even though very few words in the corpus actually express the sentiment. These extra features have two clear drawbacks. First is that they make classification slower, since there are far more words than required. The second is that they can actually reduce accuracy, since the classifier must consider these words when classifying a document. Clearly, there is an advantage in using fewer features. So, in order to remove some of the unnecessary features, feature identification is an important step. This has to be included before sentiment classification. Feature identifiation methods used commonly in opinion mining are n-grams model, bag of words model, parts of speech tagger and word relation dependency parsing trees.

Parts of Speech tagging

Parts of speech tagging (POS) also called grammatical tagging or word-category disambiguation [4,11], is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

C) Sentiment Prediction

Sentence is predicted based on the positivity and negativity of the sentence as follows.

WordNet

Sentiment words are recognized by wordnet and sentiments are assigned with corresponding scores like 1, 0 0.5 where 0 is negative sentence, 0.5 is neutral score, 1 is positive sentence.

Bayesian Network

Bayesian networks are widely used to perform classification tasks, with the following advantages.

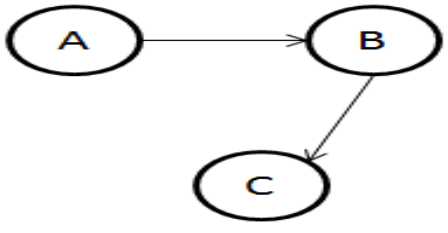
- Based on probability theory.
- Allows rich structure.
- Can mix expert opinion and data to build models.
- Backwards reasoning - in addition to predicting outputs from the given inputs, we can also use output values to infer the inputs.
- Support for missing data during learning and classification.

In most of the sentiment analysis works, Naive Bayesian classifier algorithm [6] is considered By

using the Bayesian networks the dependencies in the opinions can be summarized.

Example dependency comment: *This is not a grand laptop.*

Here "not" is a dependency word. Dependencies were not considered in the previous works and observing the sentence it looks like a negative sentence but it is a positive sentence and there is a positive word. Hence, it is a positive sentence. So, by applying Bayesian network to the above sentence dependencies are resolved by figure 2



"Figure 2: Bayesian Network Model"

Finally the opinions are ranked based on the scores like 1, 0, 0.5 where 0 is negative sentence, 0.5 is neutral score, 1 is positive sentence.

4. "Conclusion and Future scope"

This architecture solves the problems in opinion mining and provides a novel approach for sentiment classification. It also captures the opinions from various social networks and helps the customer to analyze the products and its features. It also helps the companies and organizations to improve the quality of the product based on ranking. We can analyze hundreds of comments posted by customers and rank them accordingly. An improved stemming algorithm has been implemented for better accuracy. This work can be extended by applying various machine learning algorithms with respect to the performance measures such as precision and recall for various opinions of sentiment analysis.

5. "References"

[1] Ahmed Abbasi, *"Intelligent Feature Selection for Opinion Classification",* University of Wisconsin-Milwaukee, - IEEE 2010.

[2] Bing liu, *"Sentiment Analysis: A Multifaceted Problem",* University of Illinois-Chicago, - IEEE 2010

[3] Minqing Hu and Bing Liu ,Mining Opinion Features in Customer Reviews , Department of Computer Science University of Illinois at Chicago In American Association for Artificial Intelligence 2004.

[4] http://nlp.stanford.edu/software/tagger.shtml

[5] M. Gamon, A. Aue, S. Corston-Oliver, and E. K. Ringger. Pulse: Mining customer opinions from free text.IDA'2005.

[6] Trivedi Khushboo N, Swati K. Vekariya Prof.Shailendra Mishra, Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naive Bayesian Algorithm Int.J.Computer Technology and Applications,Vol 3 (3), pages 987-991, 2012.

[7] Akshat B et al.Towards Enhanced Opinion Classification using NLP Techniques Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP) 2011

[8] http://nltk.googlecode.com/svn/trunk/doc/api/nltk.tokenizemodule.html

[9] B Jayanag et al., Feature Subsumption for Sentiment Classification of dynamic data in social networks using SCDDF, Published in International Journal of advanced Computer science and Applications vol 3 Number 9 page no 42-47, 2012.

[10] http://wordnet.princeton.edu/

[11] Dave. K., Lawrence. S., and Pennock. D., 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW'03.

[12] http://tartarus.org/martin/PorterStemmer/