A Study on Feature Subsumption for Sentiment Classification in Social Networks using Natural Language Processing

B. Jayanag
Department of Computer Science and Engineering
V. R. Siddhartha Engineering College
Vijayawada, India

K. Vineela
Department of Computer Science and Engineering
V. R. Siddhartha Engineering College
Vijayawada, India

S. Vasavi, PhD.
Department of Computer Science and Engineering
V. R. Siddhartha Engineering College
Vijayawada, India

ABSTRACT

In the past, whenever a customer wants to buy some product he used to consult his family members or friends. But this thing has been changed over the last few years where the users are generally finding out the reviews from the internet before purchasing the products. It is easy to process a review if the opinions are less, but for few popular products the reviews can be more that sometimes they will be in hundreds or thousands. It is a quite time taking process for the customers to go through all these reviews. So a system that could automatically summarize the opinions could be useful to the customers.

This paper studies existing methods for sentiment classification and proposes new method Sentiment Classification for Dynamic Data Features (SCDDF) that not only considers many sites for sentiment classification but also aggregates the opinions using Bayesian Networks and Natural Language Processing techniques. We consider the various products and their features and classify them. Bulk amount of dynamic data is considered rather than the static one. It takes as input a collection of comments from the social networks and outputs ranks to the comments in the social networks, for each product, and also classifies the comments posted. Thus the user can evaluate the product and its features.

General Terms

Sentiment Classification, Opinion Mining

Keywords

Features, feature identification, Natural language processing (NLP), opinions, Quick Test Professional (QTP), sentiment classification, summary generation, sentiment prediction.

1. INTRODUCTION

The availability of public opinion over the internet has been growing these days. It is easy to review few opinions posted over the internet. But for popular products like “IPHONE” etc the opinions may be in hundreds and thousands. It is very hard to read each and every review and also it is a time taking process. Sentiment classification on products and its features could be useful to the customers to know the reviews of previously used customers to make decisions. It is useful to the organizations and companies to know their customers feedback, so that they can enhance their products. Sentiment classification is one of the on going research fields today. Many researchers are still focusing on the methods used for this classification that could give better accuracy.

In this research, we study the previous researchers work and solve the drawbacks in their works. Our proposed architecture resolves the dependencies in the opinions and provides a summary to the customers. Feature subsumption means giving a general summary for the opinions. The proposed SCDDF consists of four phases: (1) Pre-processing the collected opinions set is preprocessed to reduce the data size, (2) features are identified on which the customers have written their opinion, (3) for each opinion the sentiment is predicted, (4) a summary is generated with the results obtained so that the customer can easily evaluate the product and get overall view of the product.

Our proposed architecture is different from that of Minqing Hu et al. [8] Feature-based opinion summarization. First of all the pre-processing step included is useful to reduce the size of the data to be processed. Secondly we consider the dependencies which are mostly neglected by most of the researchers.

As indicated our proposed architecture is divided into 4 main phases:
(1) Pre-processing- the opinions from the social networks are collected using the web crawlers and natural language processing techniques are used.
(2) Feature Identification- is the step where the features of the product are identified using parts-of-speech (POS) tagging.
(3) Sentiment prediction- predicts the overall sentiment for the review rather than for each sentence by considering the dependencies present in the review posted using the Bayesian network.
(4) The last step is Summary generation where the previous steps results are aggregated and the results are shown to the users.
The detailed study of these phases and techniques used are provided in Section 3.

2. RELATED WORK

The researchers concentrated on opinions in individual sites and also limited the data set. Many existing works are limited to the number of comments they consider. The current studies are mainly focused on mining opinions in reviews and/or classify reviews as positive or negative based on the sentiments of the reviewers.

Ahmed Abbasi [1], worked on feature selection methods and considered Intelligent Feature Selection (IFS) approach that uses syntactic and semantic information to refine larger input features to improve he opinion-classification performance. They used character n-grams, word n-grams, parts-of-speech (POS) tag n-grams, word plus POS tag n-grams, n-legomena, information extraction patterns (IEP) and semantic patterns in IFS. For each category, unigrams, bigrams, and trigrams are used. But these information modules need to be expounded on, and real-world knowledge bases could be considered. And also along with identifying the features the dependencies with in the opinionated text could have been resolved.

Andrea Esuli et al. [2], discussed on sentiment quantification i.e., whether the opinions are to be considered at individual level or aggregate level. They said that the classification is more difficult task than sentiment quantification. They used mathematical measures like Earth Mover's Distance (EMD) to evaluate sentiment quantification. But haven't described how the opinions are considered at aggregate level.

Bing liu [3], worked on opinion targets i.e., product, service, individuals, organizations or events. They said that the objects have a set of components and a set of attributes which they selectively called features. Object model a model of an opinionated text and the mining objectives and called it feature-based sentiment analysis model. They classified opinions into two types: direct opinions and comparative opinions. Direct opinions are those which have orientation of the opinions on a feature. And comparative opinions are those which compare two or more objects. Their research contains object identification, feature extraction and synonym grouping, opinion-orientation determination and integration. In object identification the objects are identified, the features like voice, sound etc are identified in feature extraction, and the opinion orientation is used for finding out the sentiments positive or negative for a given opinion. Finally they integrated these tasks by applying natural language processing techniques but the detailed description on their application is not presented and this work is not much different from the previous research works by B. Liu [10].

Claire Cardie et al. [4], has made research on fact-based question answering. They proposed an approach to multi-perspective question answering that views the task as one of opinion-oriented information extraction. They described low-level representation of opinions, and present results of interannotator agreement studies. Finally, they outlined an approach for the automatic construction of opinion-based summary representations. They created opinion-oriented “scenario templates” for summary representations of the opinions expressed in a document, or a set of documents to perform question answering. They did not identify product features and user opinions on these features to automatically produce a summary.

Hsinchu Chen et al. [7], studied opinion, sentiment, affect and subjectivity expressed in the text. Their research work is done to understand the stock performance of a large U.S corporation Wal-Mart. Their research is based on a Market Intelligence 2.0 (MI2) analysis framework. Instead of shareholder view of participants they considered stakeholder perspective. They developed a framework for analysis with 4 major stages: stakeholder analysis- for identifying the stakeholder groups participating, topical analysis- topic of discussion are determined, sentiment analysis- assessing the opinions and stock modeling- the relationships are examined. But the detailed implementation of these models is not described. And also opinions are considered only at individual level that is they collected messages from yahoo finance Wal-Mart forum only.

Minqing Hu et al. [8], 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 are also 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 SentenceOrientation procedure but using the wordnet based score and Bayesian networks the results could be better.

Most of the works in opinion mining are built around some closely related research areas such as sentiment classification, subjectivity classification, etc. Here we discuss about

(1) Sentiment Classification

(2) Subjectivity Classification

(3) Naive Bayesian Classification

2.1 Sentiment classification

These days sentiment classification has been used in opinion mining. Sentiment Classification determines the orientation of words, sentences, paragraphs and documents that is positive or negative or neutral. Dave. K et al. [5], worked on semantic classification of reviews as positive or negative or neutral ones using the available corpus from web sites, where each review already had a class e.g., binary ratings or thumbs-up and thumbs-downs. Sentiment classifiers are built around them. However, the performance was limited because a sentence contains much less information.

2.2 Subjectivity Classification

Subjective classification is different from sentiment classification. It differentiates sentences, paragraphs or documents that present opinions/comments/aspects from the factual information. Subjectivity classification aims at finding whether a sentence/paragraph/document has an opinion or not. It doesn't aim at classifying opinions as positive or negative. Sometimes subjective classification is used as a pre-processing step for sentiment classification. Wiebe. J [9] used word clustering to find the adjective features for subjectivity classification.

2.3 Naive Bayesian Classification

Naïve Bayesian algorithm [11] has been widely used for document classification, and shown to produce very good performance. A Bayesian classifier with a bag of words representation is the simplest statistical method when compared to naïve bayes classifer for resolving dependencies

$$P(\text{opinion} \mid W) = P(W \mid \text{opinion})\, P(\text{opinion})$$

$$W = \text{n-gram, POS} \wedge \text{opinion:} = \text{op}$$

Thus the existing works are limited to a particular site or a static data set. And the opinions are just classified as positive opinions and negative opinions without considering the dependencies. Naïve Bayes classifier is used for sentiment classification. But the dependencies that exist within the words in the comments are not considered. Section 3 presents our proposed system for sentiment classification.

3. PROPOSED SYSTEM

The proposed system is a unique system which takes the data dynamically, classifies, ranks are given. These ranks may vary with time and comments posted. According to the users wish the comments from the specified site are retrieved and analyzed. Comments considered here are about mobile phones and their features. Using this system the user can know the pro’s and con’s about a product. Figure 1 present the SCDDF architecture of our proposed system.

3.1. Preprocessing

3.1.1 Crawl for opinions

To get the data dynamically web crawlers such as [12] can be used. By using the web crawlers data is posted into the excel sheets. Even in excel we have data crawler to get the data from web sites, using it we can get the data into excel sheets. The tool used here is QTP. Using the recursive calls in QTP the comments posted in a site are retrieved.

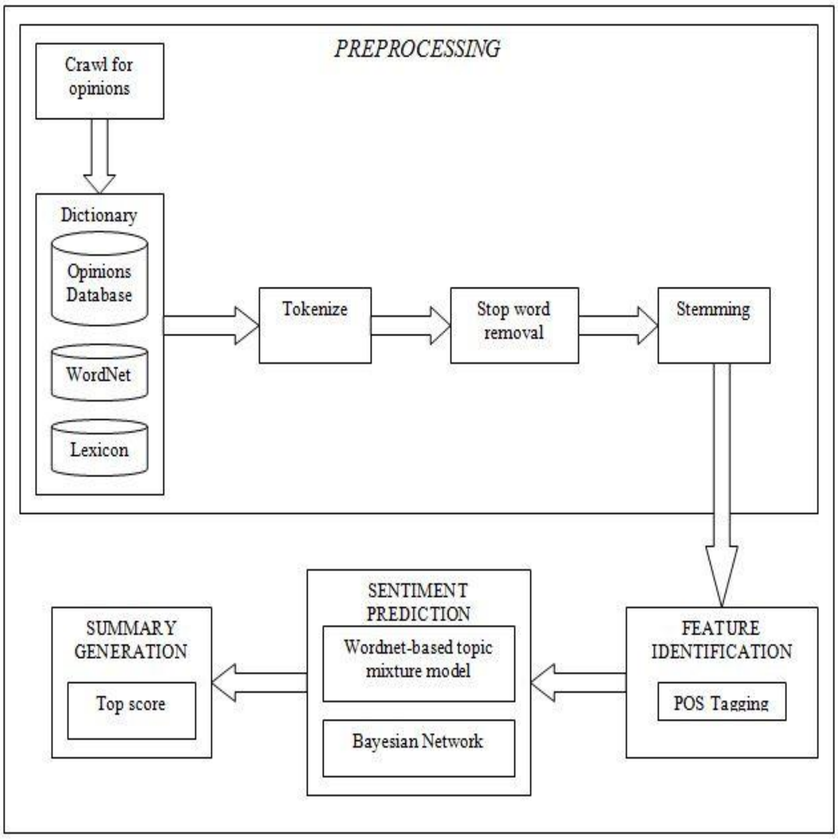

Fig. 1 Sentiment Classification for Dynamic Data Features (SCDDF).

3.1.2. Data dictionary

Data dictionary which is one of the important components of our architecture consists of opinions database, wordnet and lexicons data set.

3.1.2.1. Opinion Database

After the crawling is performed the opinions database is prepared and it is updated in intervals of time such that ups and downs in the ranks can be computed. Also the opinions that are posted are in raw form. They need to be processed.

3.1.2.2. WordNet

WordNet is a lexical reference system. WordNet resembles a thesaurus, it consists of groups words together based on their meanings. WordNet interlinks specific senses of words such that words that are found in close proximity to one another are semantically disambiguated. WordNet also labels the semantic relations among words. For this purpose we use WordNet [6, 14] for sentiment prediction.

3.1.2.3. Lexicon data set

We construct our own data set for stop word removal. For example the words like: “a”, “is”, “the”, “they”, “those”, “I”, “we”, “all”, “which”, “when”, “in”, “on”. This data set is prepared by excluding the dependency words like: “and”, “not”, “but”, “in”, “un”, etc. At present we consider 408 words corpus. Using this data set the stop words in the opinions are removed.

3.1.3. Tokenize

Tokenization is a process in which words within the given opinion are identified. We use Natural Language Toolkit Tokenizers for identifying the token.
Example for tokenizing:
The camera is good.
|The| |camera| |is| |good| |.|

3.1.4. Stop word removal

This step discards unimportant words so that the size of the data set for further phase is reduced.
Phases:

- Convert text into lower-cases
- Remove numbers and non-alphabetic symbols
- Remove punctuation (including end of sentence boundaries)
- Keep the paragraph breaks
- Remove the stop-words (with the help of lexicon)

Lexicon data set prepared by us includes all the stop words that can be removed. The stop word data sets which are available may contain the dependency words like ‘and’, ‘not’ etc., which are to be considered to solve the dependency problems, so we need to prepare the data set.
Example:

The battery is good.
battery good
Stop words here are: “The”,”is”,”.”
Also some words which are considered as stop words in Natural Language Processing are important in sentiment classification such as “not”.

3.1.5. Stemming

This is a step of finding morphological number of words as given below*:*
COLLECT
COLLECTS
COLLECTED
COLLECTING
The suffix stripping process will reduce the total number of terms and hence reduces the size and complexity of the data in the system, which is always advantageous. The porter stemmer algorithm is used for suffix stripping.
Porter stemmer[13] is a 6 steps algorithm, where in each step the words are trimmed and the size of the data set will be reduced in each step. The suffix stripping process reduces the size of the vocabulary by about one third. Thus by using this algorithm we can reduce the number of words in the given comment and reduce the data size.

3.1.6. Cryptics

The full forms of certain words are expanded in this step.
Example: id- identifier
TV- television
avg- average

3.1.7. Intermediate Databases

3.1.7.1. Updated Database

A reduced and meaningful database is constructed in the pre-processing step after the tokenizing**,** stop word removal and stemming. And this database is sent to remove the spam data.

3.1.7.2. Product Database

Spam data is not necessary for sentiment analysis so in the pre-processing step this spam is removed by the help of product data base. Here we construct a data base and we compare the updated database with the product database and we remove all the ambiguous or spam data.

3.1.7.3 Processed Database

Finally after the pre-processing is done a new data set is created for sentiment analysis. It doesn’t contain stop words or spam data.

3.2 Feature Identification

To analyze the data first the features must be identified. Features are like battery, camera, sound, touch, memory, etc. Here we try to identify the nouns, noun phrases, verbs and verb phrases. This can be done by using NLP techniques such as parts-of-speech (POS) tagging, n-grams, edit distance etc. Our method considers POS tagging.

3.2.1. POS tagging

It is used in the identification of words as nouns, verbs, adjectives, adverbs, etc. Each token in the given sentence is tagged with its respective parts-of-speech (POS) using the Wordnet. Thus the tokens/words are tagged with their respective parts-of-speech.

3.3. Sentiment Prediction

Sentiment prediction is used to know the positivity and negativity as follows:

3.3.1. WordNet-based topic mixture model

Sentiment words are identified by using wordnet and each sentiment word will be assigned with the corresponding positive or negative or neutral score. For example excellent will be assigned a score 1, worst 0, average 0.5 etc. based on the scores assigned finally mixing all the scores the score for the product is given.

3.3.2. Bayesian Network

In most of the sentiment analysis works Naive Bayes classifier model is considered but the Naive Bayes model doesn’t solve the dependency problems in the opinions. By using the Bayesian networks the dependencies in the opinions can be summarized.
Example:
The camera is not bad.
The camera is bad.
The camera is good.
Sentences 2 and 3 are direct sentences with out any dependency, sentence 2 is a negative opinion and sentence 3 is a positive opinion. When sentence 1 is considered the prior sentiment analysis methods say that it is a negative opinion because there is a negative word in it.
But here it is a positive sentiment; if we don’t consider the dependency as a negative word is present it says that the sentiment is a negative opinion. We can avoid this by considering the conditional independence model in Bayesian networks.
For each and every comment dependency word count is first initialized to 0 and when a dependency word appears and it is followed by a sentiment word then the dependency word count is turned to 1. And then the bayesian network model is applied to the comment.
Let us consider a comment and pass it through the bayesian network. “This phone is not bad”. After applying the above techniques all the stop words will be removed and the POS tagged main words “phone” “not” “bad” will be left. When this comment is considered in sentiment prediction level as a dependency word “not” is present, the dependency word count will be updated to 1. And the word is immediately followed by a sentiment word “bad” so the opinion will be passed through the bayesian network. Now 1-propability of that word will be performed and the resultant value is given.

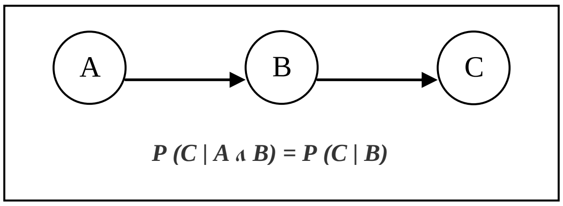

Fig. 2 Bayesian Network model

Here let’s assume A is not, B is bad. Thus “not bad, which means good” is the result.

3.4. Summary Generation:

Finally the results from sentiment prediction are considered and summary is generated. The positive and negative counts generated for each feature are shown and based on this count the overall positive and negative count is given for the product. By using this score the Statistical Summary is shown to the user, thus user can easily understand the pros and cons of the product. Methods used in summary generation are: Statistical summary, aggregate rating, summary with in a time line and our proposed method uses statistical summary.

4. CONCLUSION

An architecture that improves the sentiment classification considering the dependencies was proposed. It solves the problems in opinion mining and provides a novel approach for sentiment classification. It is a novel architecture that successfully captures the peculiarities of social networks. Sentiment analysis on product features is useful for customers and shoppers to know the pros and cons of the product. It is also useful for companies and organizations. They can use the architecture to enhance their products. Furthermore, we note that dependencies are significant features for sentiment classification and are most important than the spam data.

Our approach provides the specific knowledge that can help the users to make the right decisions. In this architecture we considered techniques that could improve the classification performance. However, the success of such an initiative eventually depends on the cooperation of the companies and institutions owning social network data, and on the agreement of enough organizations to participate in such a project.

5. FUTURE WORK

Our proposed method has to be evaluated on the data set available in the social networks and performance in terms of precision, recall and time taken has to be compared with existing models. Also we should check how far ontology's help in enhancing results of sentiment analysis.

6. REFERENCES

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

[2] Andrea Esuli and Fabrizio Sebastiani, *"Sentiment Quantification",* Italian National Council of Research.

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

[4] Cardie, C., Wiebe, J., Wilson, T. and Litman, D. 2003. "*Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering".* AAAI Spring Symposium on New Directions in Question Answering. 2003.

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

[6] Fellbaum, C. *"WordNet: an Electronic Lexical Database"*, MIT Press 1998.

[7] Hsinchu Chen and David Zimbra, "*AI and Opinion Mining*", University of Arizona, - IEEE 2010

[8] Minqing Hu and Bing Liu, "*Mining and Summarizing Customer Reviews",* KDD 2004: 168-177

[9] Wiebe, J. 2000. *"Learning subjective adjectives from corpora".* In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence. AAAI Press, 735–740.

[10] B. Liu, "*Sentiment Analysis and Subjectivity,"* Handbook of Natural Language Processing, 2nd ed., N. Indurkhya and F.J. Damerau, eds., Chapman & Hall, 2010, pp. 627–666.

[11] Pop, I. *"An approach of the Naïve Bayes Classifier for the document classification",* General Mathematics, Vol. 14, No.4, pp. 135-138, 2006.

[12] Jeff Heaton, *"Programming Spiders, Bots, and Aggregators in Java",* Publisher: Sybex, February 2002, ISBN: 0782140408

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

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