A Feature Based Approach for Sentiment Analysis using SVM and Coreference Resolution

Hari Krishna M, Rahamathulla *K*
Department of Computer Science and Engineering
Govt. Engineering College
Thrissur, India

Ali Akbar
Department of Computer Science and Engineering
Govt. Engineering College
Wayanad, India

***Abstract*—Online shopping is one of the most comfortable ways to shop in this new era of technology. People buy online products frequently and post their reviews about the products they have used. The viewpoint of the user will be in the form of tweets or product reviews which they post in an e-commerce site. These reviews will have significant role in deciding how far the products have been placed in peoples mind. These reviews will also help the manufacturers to improve the features of the product as required But it is a very difficult task to manually read the reviews and assign sentiment to them. This problem can be solved by creating an automated system in which we can analyze the reviews posted by the users and extract the users perception about a particular feature. In this work we have developed a procedure for 'feature based sentiment analysis' by using a classifier called Support Vector Machine.**

Keywords—sentiment analysis; natural language; data mining; Stanford parser; product reviews.

I. INTRODUCTION

Nowadays enormous amount of user opinionated data flood the Internet. The purchaser can form an opinion about the product they intend to buy from the product reviews in e-commerce site, blogs etc. Since massive amount of opinionated data flows in to the internet it is pretty difficult for the user to examine them and form an opinion. So we need to have an automated approach. The machine can be trained to analyze the text written by a reviewer and identify whether it is positive or negative. This is an efficient method because we can process a lot of user opinion in very little time [1].

In this research we are using a machine learning approach called supervised classification. This method tends to be more accurate than other methods as we are training the classifier using real world dataset. There should be two types of datasets. One is called training set and it is created manually. They are basically used to train the classifier. We use test set to determine the accuracy of the system. There are different types of machine learning methods which can be used for classification and they are Naive bayes, Maximum Entropy(ME) and SVM.

In this work we have used SVM as the classifier. At the end we detect the sentiment users have towards each feature of a product. We also detected the overall sentiment about the product. Negations were considered manually while creating the data set and the problem is resolved.

The rest of this paper is organized as follows: Section II presents the literature survey; Section III presents the system architecture and details of the proposed system and we conclude this paper in Section IV .

II. LITERATURE SURVEY

Nagarjuna et al.[2] introduced a method to identify the sentiment about each feature of a product. This approach does not identify the anaphoras in a review. To overcome this problem we have introduced a coreference resolution module in our system.

In [3], Christopher et al. proposed a supervised machine learning method to classify web reviews. Classifying web reviews are quite difficult because texts in web reviews are quite less rigorous than those in formal documents such as newspaper articles, business reports and journal articles.

Sentiment classification and analysis are the two problems in opinion mining. Sentiment classification is used to identify the sentiment of a review irrespective of which product feature it is describing. The basic task of sentiment analysis is to classify review opinions into number of product feature classes.

In [4], Subhabrata et al. proposed a novel approach to identify the features present in a review and extract opinion expressions about those features by exploiting their association. Later these opinion expressions are classified as positive or negative. For feature extraction they propose two methods. The first method is used for those reviews whose domain is not known and the other method is for those reviews whose domain is known. After obtaining the features the

978-1-5090-5297-4/17/$31.00 ©2017 IEEE

relation between opinion expression are obtained using Stanford dependency parser.

III. Proposed System

The proposed system is an improvement to the technique introduced in [2]. Fig. 1 shows the proposed system architecture.

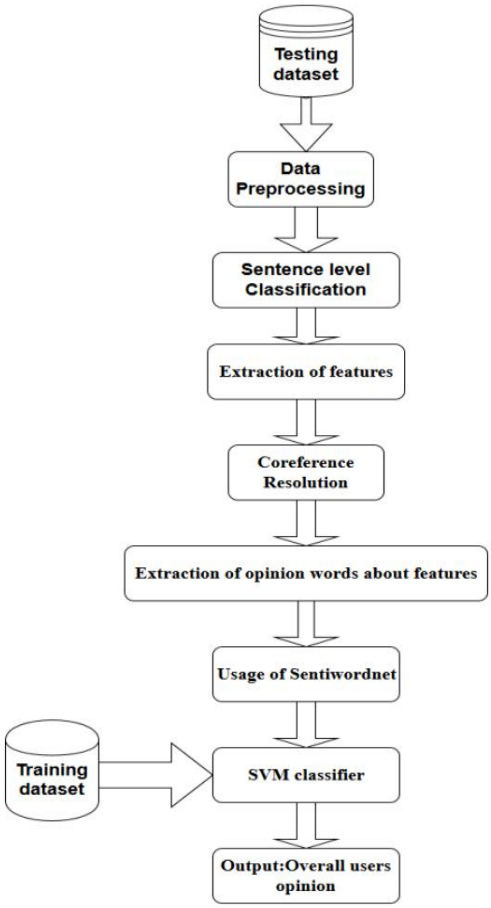

Fig.1. Proposed System Architecture

A. Dataset

Here two types of datasets are created .One is for training and other is for testing. Both are created manually.The training set is basically a X:Y relation in which X is score of a probable opinion word and Y represents whether that score is positive or negative. Testing set is created by taking reviews from e-commerce sites. Each review in the testing set are manually tagged whether they are positive or negative. After training is completed, system will be able to comprehend which reviews will have positive sentiment and which will have negative sentiment. We intend to test the system by giving reviews from test set whose polarity is already known to us. Depending on the output produced by the system based on testing set we can determine the accuracy of the system.

B. Data Preprocessing

The main preprocessing methods performed are stemming, error correction and stop word removal [12]. Stemming is the basic task of identifying the root of a word. The purpose of this method is to remove various suffixes, to reduce the number of words and to save time and memory. We need to incorporate error correction as the reviewers will not be following the exact grammatical rules, punctuation and spellings. These mistakes make the system to understand the context in a different way and need to be corrected. Stop words are removed to make the text less complex. Some stop words like “it” should not be removed as it may effect coreference resolution.

C. Sentence Level Classification

Every sentence which has a positive, negative sentiment is called subjective sentences. Some users though may ask questions or write sentences in reviews which does not convey any sentiment. These sentences are called objective sentences. We eliminate all such sentences with the aim to reduce the overall size of the review. Sentences which contains words like where and who has a high chance of becoming a question. A question doesn’t convey any sentiment and need to be removed. Questions can be identified using regular expression in python [12].

D. Extraction of Features

It’s one of the most difficult problems in Sentiment analysis. The features of a product would always be a noun. So to identify the features we need to identify and extract all the nouns in a review using POS-tagging. Features which rarely occur need to be removed. After removal of such rare features we would get a list of features which occur prominently. Frequent features of a mobile phone would be its camera, display, storage, processor, memory and so on.

E. Co-reference Resolution

Main aim of this module [5][6][7] is to remove anaphora’s occurring in a review. We can remove them using coreference resolution. In this system the opinions about a particular feature are obtained using a sentence level search. For example consider two sentences which occurred in a review, “Battery of this phone is 3900 mAh. It is awesome”. In this example we can clearly see that the word awesome is about the phone feature Battery, but we are not able to detect that as ”awesome” is in another sentence. When coreference resolution is used the word “it” gets replaced by the feature name. In this case “it” gets replaced by “Battery”. So we will be able to extract the word awesome for the feature Battery. We have used Stanford deterministic coreference resolution system.

F. Extraction of Opinion Words

Words which convey a sentiment about a particular feature can be obtained using Stanford parser [10][11]. The parser will give a collection of grammatical dependencies between words in a sentence as its output. We have to look at the

978-1-5090-5297-4/17/$31.00 ©2017 IEEE

dependencies to find the opinion word for the features which we have already extracted in the previous step. We notice that we can directly find the opinion words for some features and such type of dependencies are called direct dependencies. Apart from direct dependencies we also need to consider transitive dependencies.

G .SentiWordNet

The Sentiwordnet is especially created for opinion mining applications [9]. For every word in SentiWordNet there are 3 polarities associated with it .The 3 polarities are positivity, negativity and subjectivity. For example the score of the word high in SentiWordNet is .125, but sometimes in sentences like "cost is high" the word high cannot be considered as positive. In turn it conveys a negative meaning. So we have to consider such situations.

H. Support Vector Machine

SVM is chosen as the classifier [8]. We use SVM because sentiment analysis is a binary classification and also it has the capability to work with huge datasets. In this case, in order to train the classifier we have used a manually created training set. The training set is basically a X:Y relation which x is score of a opinion word and y is whether that score is positive or negative. The input we gave to SVM is basically a score of the opinion word about a feature in a review.

I. Extraction of Feature Wise Opinion

To extract the opinion about a specific feature, we need to consider all the reviews which contain that feature. The eventual positive score for specific feature is the ratio of total number of reviews in which there is a positive sentiment about the feature to total number of reviews. The eventual negative score for specific feature is the ratio of total number of reviews in which there is a negative sentiment about the feature to total number of reviews.

J. Extraction of Whole Product Opinion

To discover the overall opinion, reviewers have about a product we have to consider the entire reviews we have used to test the system. For every review, we separately calculated a sum of their positive and negative score and we finally have one positive and negative score for each review. Now we compute the total number of reviews whose positive score is greater than the negative score and these are called positive reviews and rest of the reviews are called negative reviews.

IV.Conclusion

In this paper we have proposed a combination of SVM and coreference resolution to improve the accuracy of feature based sentiment analysis. Stanford dependency parser has been used to extract opinion words about the features and SentiWordNet has been used to assign scores to these opinion words.

Future research can be focused on identifying reviews which are sarcastic. It is very difficult for both humans and computer to detect. Another issue we can focus on is spam review identification. Sometimes the reviewers may post advertisement in the review section. It needs to be detected and discarded.

References

[1] Bing Liu, Sentiment Analysis and Subjectivity, Handbook of Natural Language Processing, 2010.

[2] D V Nagarjuna Devi, Chinta Kishore Kumar and Siriki Prasad , "Feature Based Approach for Sentiment Analysis by Using Support Vector Machine",IEEE 6th International Conference on Advanced Computing, 2016.

[3] Christopher C. Yang ,Y.C. Wong, Chih-Ping Wei,"Classifying Web Review Opinions for Consumer Product Analysis", ICEC '09, August 12-15, 2009, Taipei, Taiwan

[4] Subhabrata Mukherjee, Pushpak Bhattacharyya,"Feature Specific Sentiment Analysis for Product Reviews", 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I

[5] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky, "Deterministic coreference resolution based on entity-centric, precisionranked rules",Computational Linguistics 39(4), 2013.

[6] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu,and Dan Jurafsky, "Stanfords Multi Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task", In Proceedings of the CoNLL-2011 Shared Task, 2011.

[7] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, Christopher Manning, "A Multi-Pass Sieve for Coreference Resolution EMNLP-2010", Boston, USA, 2010.

[8] Martín-Valdivia M T, Rushdi Saleh M, Ureña-López L A, Montejo-Ráez A, Experiments with SVM to classify opinions in different domains, Expert Systems with Applications, 38(12), 14799- 14804, 2011.

[9] Andrea Esuli and Fabrizio Sebastiani, Sentiwordnet: A Publicly Available Lexical Resource for Opinion Mining, in LREC 2006.

[10] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In 5th International Conference on Language Resources and Evaluation (LREC 2006).

[11] Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In COLING 2008 Workshop on Cross-framework and Crossdomain Parser Evaluation.

[12] Steven bird,Ewan klein, Edward looper"Natural language processing with python"

978-1-5090-5297-4/17/$31.00 ©2017 IEEE