IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Reversed Sentiment Analysis on Product reviews

Mukta Raut[1], Mayura Kulkarni[2], Dr. Sunita Barve[3]

Student, Department of Computer Engineering, MIT Academy of Engineering, Pune, India[1]

Sr. Assistant Professor Department of Computer Engineering, MIT Academy of Engineering, Pune, India[2]

Associate Professor Department of Computer Engineering, MIT Academy of Engineering, Pune, India[3]

Abstract: The Web has rapidly changed the way of gathering opinions about certain assets from word-of-mouth to browsing online feedbacks. In doing so, one has to go through a huge amount of opiniated data available on internet and analyze it to get the user reviews about a particular product or service. It becomes rather tedious for a human to manually go through each and every review in order to evaluate the product, as the volume of such data available online is huge and scattered across different sources such as websites, blogs, twitter, etc. The need for an automated system to analyze such data and generate reliable results is one of the root causes of upsurge in research on sentiment analysis and opinion mining. In this paper we propose a system which evaluates any product or service based on sentiment analysis of the online reviews. We use a dual sentiment analysis method to classify reviews into three classes- positive, negative and neutral. Classifier used is Naïve Bayes classifier.

Keywords: Natural language processing, Machine Learning, Sentiment Analysis, Opinion Mining

I. INTRODUCTION

Online shopping, socializing, marketing of products and services, online education are increasing rapidly with corresponding growth in internet availability and speed. E-commerce is relatively new for customers who are used to shop directly in retail shops and shopping malls. Thus in case of online shopping customer satisfaction is crucial as this industry is in development phase. So the customer feedback is most valuable to analyze the business success. This has increased the importance of sentiment analysis of the consumer reviews, which is one of the core areas of research in data mining and information retrieval.

Huge amount of opiniated data is created each hour in forms of reviews, feedbacks, online blogs, social networking sites and so on. This data is mostly unstructured and comes in different types such as star ratings, numerical rating on 1 to 10 scale, text reviews and so on. To analyze this data accurately for its sentiment interpretation, the easiest way is to go through it manually. Although this method can give most accurate results, it is practically impossible to manually collect reviews from hundred different sources and analyze them one by one. To do this job efficiently, we need automated systems which can perform sentient analysis on opiniated data. Such system makes use of machine learning techniques, statistics and natural language processing to extract, identify, or otherwise characterize the sentiment content of a text unit. These tasks are collectively known as sentiment analysis or opinion mining. The automated techniques developed for sentiment analysis saves the overhead of manual analysis of the big data. The main advantage of sentiment analysis is speeding up the decision making process of the consumers without compromising on time and the quality of product evaluation. To producers, sentiment analysis benefits by using this evaluation for maintaining the service quality up to mark and predicting the upcoming tends in consumer markets. The other applications of opinion mining are studying the changes in customer psychology, analyzing public opinion about a certain event, detecting aggressive, provoking communication patterns over the social communication sites (e.g. detection
of any movement of anti-socialite bodies in sensitive areas), predicting results for elections, Public Opinion Polls, etc. There are many academic and commercial tools available to perform sentiment analysis such as Oracle (www.oracle.com/social), IBM(www.IBM.com/analytics), SenticNet (www.business.sentic.net), Luminoso (www.luminoso.com), etc. They give the graphical summarization of the public opinions collected from websites on a very large scale. The main problem in using such high level tools for natural language processing is that they are limited to particular domain. In this paper we present a simple technique which classify the directional text in form of reviews into positive, negative and neutral class.

II. PROBLEM DEFINATION

In many cases, opinions are hidden in long forum posts and blogs. It becomes tedious for a human analyst to find relational sources, extract relational sentences with opinions, read them, understand them, and organize them into usable form. Thus, automated summarization systems are needed. We can use the results of these systems for sentiment analysis applications which include varied fields such as social media, e-commerce, online service portals, online

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

survey and feedbacks, product launch, promotion and advertising, etc.This proposed system targets reviews and feedbacks about hotels from online hotel booking sites. This System is specifically made for- when a user wants to use a certain online service, his choice of service provider is based on the feedback of consumers which had been using the service before. The system collects such feedback into a database and gives results in the form of pie chart, showing positive and negative aspects of the service provider, and rates the system accordingly. So consumer can simply look at the results and make choice, without having to go through various feedback sources for different service providers.
In our system we address two problems of such systems: polarity shift problem and domain-specific system.

III. LITERATURE SURVEY

In [1], Erik Cambia states various advancement in affective computing and sentiment analysis and also lists all other research fields emerging from it, along with its applications. [1] also lists and explains the basic tasks in sentiment analysis namely emotion recognition and polarity detection. It also presents various sentiment detection and classification models. The existing approaches are divided into three main categories: knowledge based techniques, statistical methods and hybrid approaches. [2] illustrates a method to develop corpora for detecting and analyzing sentiments in social media with help of an Italian project senti-TUT that investigates sentiments and irony in online political discussions. In linguistics, corpus refers to a structure of words or sentences representing certain properties and used for lexical, grammatical or other linguistic analysis. Corpus development in OMSA is a three step method which includes collection, annotations and analysis each of which is strongly dependent on the other [2]. Collection consists of choosing a dataset to compose the corpus and collecting the methodologies that can be applied to it. The annotation step includes defining scheme and applying it to the collected data. Annotation helps in utilization of unstructured data for machine learning. Analysis of the developed corpus is then carried out. The annotated data is used as training and test dataset in statistical machine learning tools for sentiment classification. [3] presents the survey of related work on feature selection methods in OMSA. It presents the future scope in development of feature selection methods in OMSA. [3] also propose a rule-based multivariate text feature selection method called Feature Relation Network (FRN) that considers semantic information and also leverages the syntactic relationships between n-gram features. [4] propose learning sentiment-specific word embeddings. Word embedding are typically studied from unannotated plain text. Word embedding provides a dense vector representation of the semantic aspects of the words. However, these representations are not able to distinguish between aspects such as sentiment polarity of the text. As a result, the existing word embedding learning algorithms ignore the sentiment of texts
Thus, when we consider literature survey there are many drawbacks from existing systems which we address in our proposed system. They are listed as follows:
1. Polarity shift problem: In this when sentence contain positive as well as negative sentiments than system will confuse to give result or give wrong result.
2. Dual sentiment analysis: This system will give positive and negative sentiments.
3. Dual Expansion and Dual Training algorithm for reverse review: Dependency on external dictionary for review reversion.
4. BOW Problems: the performance of bag of words sometimes remains limited due to some fundamental deficiencies in handling the more complex polarity shift patterns such as transitional, subjunctive and sentiment inconsistent sentences in creating reversed reviews.

IV. SYSTEM METHODOLOGY

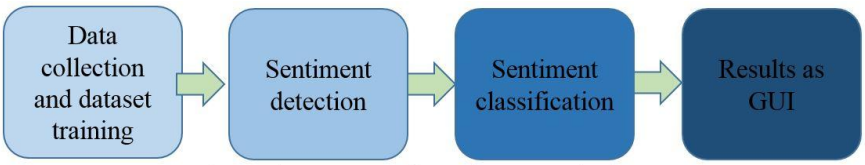

Fig. 1 The process of Sentiment Analysis.

Sentiment analysis is carried out in the following steps:
1. Data Collection and dataset training: Sentiments are usually expressed via public forums like the blogs, discussion boards, online selling websites, etc. Data collection involves extracting this data from different sources into a single data structure. This data needs to be 'cleaned' for redundancy, consistency, integration and other such properties as per requirement. This stage is also called as text preparation. To obtain trained dataset, we manually classify these reviews into positive, negative and neutral class and label them likewise.
2. Sentiment Detection: At this stage, each sentence of the review and opinion is checked for negators and sentiment reflecting words. All the negative words are removed, the sentimental words within the scope of negators are reversed.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

By this, we obtain a reversed review of the original review. Thus, we get two training datasets: original training dataset and reversed training dataset. To obtained reversed dataset, we utilize pseudo-antonym dictionary

3. Sentiment Classification: Sentiments are classified into three groups, positive, negative and neutral. At this stage of sentiment analysis methodology, each subjective sentence detected is classified into one of the above groups. Classifier used to train the dataset is Naïve Bayes classifier.

4. Product Evaluation: The main idea of sentiment analysis is to convert unstructured text into meaningful information. After the completion of analysis, the text results are displayed on pie chart.

V. PROPOSED SYSTEM

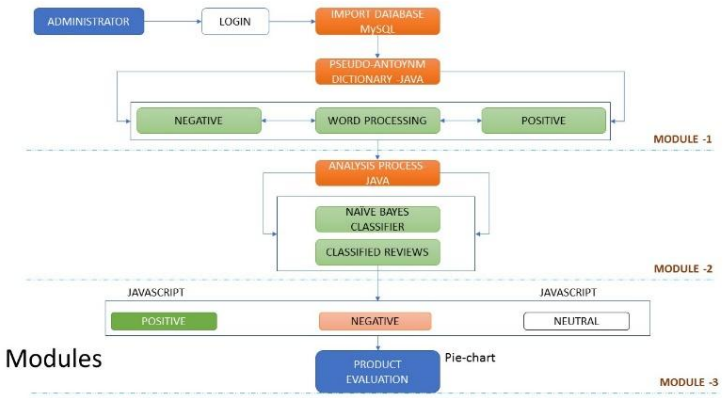

Fig. 2 Different modules of proposed system

This section gives the overview of the different modules in the proposed system. The system is divided into three main modules. Figure 2 represents different modules of system.

A. SYSTEM MODULES

Module 1: This module consists of user login, importing training dataset and word processing unit. The training dataset contains pre-classified and labeled reviews with 3-class classification. In order o train our dataset, we perform this classification manually. Then review-reversion is performed for which we use the pseudo-antonym dictionary. This dictionary contains a list of words paired with their sentiment-opposite words. We also maintain a separate list of negation words. Word processing unit uses these two lists and obtain a pair consisting of original review and opinion-reversed review. This output is then fed to second module.

Module 2: It is sentiment analysis module. For each of the original review and opinion reversed review, we calculate the 'score' using Naïve Bayes classifier. This score is then used to classify the given review into a 3-class classification i.e. positive, neutral and negative.

Module 3: Classified reviews from module 2 are input into the product evaluation unit. Using this information, we draw a pie-chart for each product to be evaluated. Thus, users can simply visualize he results at a glance through the pie-chart and make a choice.

B. MATHEMATICAL MODULE

We use naïve bayes classifier for finding 'score' of reviews. This classifier works as follows. Naive Bayes classifiers are for studying the classification task from a Statistical point of view. The starting point is that the probability of a class C is given by the posterior probability P(C|D) given a training document D. Here D refers to all of the text in the entire training set. It is given by D =(d1, d2,…di), where di is the ith attribute (word) of document D.

$$P(C = c_i|D) = \frac{P(D|C = c_i).P(C = c_i)}{P(D)}$$

P-is prior probability
$C = c_i$-is for class label
D-is for document features (Words or chunks of texts in the documents.)
$P(C = c_i|D)$ −Posterior Probability of a certain class give set of words.
(What is the class that our document is belong to)
$P(D|C = c_i)$ −Likelihood Probability of set of words given the classes.
$P(C = c_i)$-Class Prior Probability
$P(D)$ −Predictor Prior Probability

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Since the marginal probability P(D) is equal for all classes, it can be disregarded and the equation becomes:

$$P(C = c_i|D) = P(D|C = c_i).P(C = c_i)$$

The document D belongs to the class C which maximizes this probability, so:

$C_{NB} =$ argmax P(D|C).P(C)

$C_{NB} =$ argmax $P(d_1, d_{2,...,}d_{n,}|C)$. P(C)

Assuming conditional independence of the words d_i, this equation simplifies to:

$C_{NB} =$ argmax $P(d_1|C)$.P(C). $P(d_2|C)$.,.cdot $P(d_n|C)$

$C_{NB} =$ argmax P(C). $\Pi_i P(d_i|C)$

Here $P(d_i|C)$ is the conditional probability that word i belongs to class C. For the purpose of text classification, this probability can simply be calculated by calculating the frequency of word i in class C relative to the total number of words in class C

$$\mathbf{P(d_i|C)=\frac{count\ (d_i|C)}{\Sigma_i count\ (d_i|C)}}$$

We need to multiply the class probability with all of the prior-probabilities of the individual words belonging to that class. The question then is; how do we know what the prior-probabilities of the words are? Here we need to remember that this is a supervised machine learning algorithm: we can estimate the prior-probabilities with a training set with documents that are already labeled with their classes. With this training set we can train the model and obtain values for the prior probabilities. This trained model can then be used for classifying unlabeled documents.

This is relatively easy to understand with an example. Let's say we have counted the number of words in a set of labeled training documents. In this set each text document has been labeled as either Positive, Neutral or as Negative. The result will then look like:

TABLE I WORKING OF NAÏVE BAYES CLASSIFIER

Word	Positive Class	Neutral Class	Negative Class	Total
Bad	10	20	70	100
Worst	70	20	10	100
Awesome	50	200	50	600
This	100	600	100	800
Hotel	10	90	10	100
Total	240	1230	240	1700

From this table we can already deduce each of the class probabilities:

$P(C_{pos})=0.141$,

$P(C_{neu})=0.723$,

$P(C_{neg})=0.141$.

If we look at the sentence "**This Hotel is awesome**.", then the probabilities for this sentence belonging to a specific class are:

$P(C_{pos})=0.141 \cdot 50/240 \cdot 10/240 \cdot 100/240 \cdot 70/240 = 1.49 \cdot 10\text{-}^{3}$

$P(C_{neu})=0.723 \cdot 500/1230 \cdot 90/1230 \cdot 600/1230 \cdot 20/1230 = 6.82 \cdot 10^{-4}$

$P(C_{neg})=0.141 \cdot 50/240 \cdot 10/240 \cdot 100/240 \cdot 10/240 = 2.12 \cdot 10\text{-}^{4}$

This sentence can thus be classified in the positive category.

VI. DATASET DESCRIPTION

There are many datasets such as SentiWordNet, available for English language. These datasets are available easily foe languages in which lexical resources are abundant, such as English. These resources group words based on semantic similarities. Also many antonym dictionaries are available. However, these datasets are a general grouping of words and may not reflect a particular domain of our choice. Also such resources are not readily available for many of the local languages, except a few like English, French, etc.

So in our system, we manually select the reviews from across different online review sites and classify them in a 3-class classification. We also propose our own antonym dictionary with entries relevant to our domain. By doing this, we go further towards our goal of a domain-adaptive, language independent platform for sentiment classification. We use dataset in which 2000 reviews are collected from different sites and stored in MySQL database.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

VII. OBSERVATIONS

The accuracy of the review classification depends upon the class label given to training reviews accurately. We notice that in our proposed approach, sometime our algorithm is affected by the misclassified class labels in training dataset. Review reversion is affected by presence of duplicate entries in the pseudo-antonym dictionary. Also the accuracy of the system increases with larger training dataset.

VIII. CONCLUSION AND FUTURE SCOPE

Although the idea of fully automated machine leaning system seems very attractive, it is extremely difficult to train such a system to deliver desired accuracy. This is because, in human world, there is often a different interpretation of a single word. This interpretation depends on context, situation, emotional state of the person, social background, and many such parameters. So in practical applications, to provide a completely automatedsolution is nowhere in sight. However, we can devise effective semi-automated solutions. The key is to fully understand the whole range of issues and pitfalls, cleverly manage them, and determine the modules that can be fully automated and modules that need human assistance. By this we can push more and more toward automation.
In our research attempt, we have focused on the polarity shift problem type negation, which occurs most frequently. However, there two more types of polarity shifts-contrast and sentiment inconsistency. Thus, our future work would be to extend system to include above two types of polarity shifts as well.

ACKNOWLEDGMENT

It is our immense pleasure to work on this project. It is only the blessing of our divine guides Prof. Mayura Kulkarni and Co-guide Prof. Sunita Barve who has prompted and mentally equipped us to undergo the study of this project. We are also grateful to our entire staff of Computer Engineering Department, Family and Friends for their kind co-operation which helped us in successful implementation of project.

REFERENCES

[1] Affective Computing and Sentiment Analysis, Editor: Erik cambria, Nanyang Technological University, Singapore, cambria@ ntu.edu.sg, Published by the IEEE Computer Society
[2] Developing Corpora for Sentiment Analysis: The Case of Irony and Senti- TUT Cristina Bosco, University of Torino, Viviana Patti, University of Torino, Andrea Bolioli, CELI srl, Published by the IEEE ComputerSociety, 2013.
[3] Selecting Attributes for Sentiment Classification Using Feature Relation Networks Ahmed Abbasi, Member, IEEE, Stephen France, Member, IEEE,Zhu Zhang, and Hsinchun Chen, Fellow, IEEE. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011.
[4] Sentiment Embeddings with Applications to Sentiment Analysis Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 2, FEBRUARY 2016.
[5] Dual Sentiment Analysis: Considering Two Sides of One Review Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 8, AUGUST 2015.
[6] Scope of Negation Detection in Sentiment Analysis. Maral Dadwar, Claudia Huff. Human Media Interaction Group University of Twente Enschede, Netherlands.
[7] Sentiment Analysis using Product Review Data. Xing Fang and Justin Zhan. Fang and Zhan Journal of Big Data (2015) 2:5
[8] Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1-2):1135
[9] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.
[10] B. Liu, Sentiment Analysis and Opinion Mining (series Synthesis Lectures on Human Language Technologies). vol. 16. San Mateo, CA, USA: Morgan, 2012.