Sentiment Analysis on Social Media and Online Review

Rajni Singh
PG Student
Department Of Computer Science and Engineering
Lovely Professional University, Punjab

Rajdeep Kaur
Assistant Professor
Department Of Computer Science and Engineering
Lovely Professional University, Punjab

ABSTRACT

Social media monitoring has been growing day by day so analyzing of social data plays an important role in knowing customer behavior. So we are analyzing Social data such as Twitter Tweets using sentiment analysis which checks the attitude of User review on movies. This paper develops a combined dictionary based on social media keywords and online review and also find hidden relationship pattern from these keyword.

Keywords

Preprocessing, Tokenization, Removal of stop words, Stemming.

1. INTRODUCTION

Nowadays, Social media is becoming more and more popular since mobile devices can access social network easily from anywhere. Therefore, Social media is becoming an important topic for research in many fields. As number of people using social network are growing day by day, to communicate with their peers so that they can share their personal feeling everyday and views are created on large scale. Social Media Monitoring or tracking is most important topic in today's current scenario. In today many companies have been using Social Media Marketing to advertise their products or brands, so it becomes essential for them that they can be able to calculate the success and usefulness of each product [2]. For Constructing a Social Media Monitoring, various tool has been required which involves two components: one to evaluate how many user of their brand are attracted due to their promotion and second to find out what people thinks about the particular brand.

To evaluate the opinion of the users is not as easy as it seems to all users. For evaluating their attitude may requires to perform Sentiment Analysis, which is defined as to identify the polarity of customer behavior, the subjective and the emotions of particular document or sentence. To process this we need Machine Learning and Natural Language Processing methods and this is place where most of the developers facing difficulty when they are trying to form their own tools.

Over the recent years, an emerging interest has been occurred in supporting social media analysis for advertising, opinion analysis and understanding community cohesion. Social media data adapts to many of the classifications attributed for "big-data" – i.e. volume, velocity and variety.

Analysis of Social media needs to be undertaken over large volumes of data in an efficient and timely manner. Analysing the media content has been centralized in social sciences, due to the key role that the social media plays in modelling public opinion. This type of analysis typically on the preliminary coding of the text being examined, a step that involves reading and annotating the text and that limits the sizes of the data that can be analysed.

1.1Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to identify and extract one-sided information in source materials or simply it refers to the process of detecting the polarity of the text. It also referred as opinion mining, as it derives the opinion, or the attitude of a user. A common approach of using this is described how people think about a particular topic. Sentiment analysis helps in determining the thoughts of a speaker or a writer with respect to some subject matter or the overall contextual polarity of a document. The attitude may be his or her decision or estimate, the emotional state of the user while writing.

Sentiment Analysis can be used to determine sentiment on a variety of level. It will score the entire document as positive or negative, and it will also score the reaction of individual words or phrases in the document. Sentiment Analysis can track a particular topic, many companies use it to track or observe their products, services or status in general. For example, if someone is attacking your brand on social media, sentiment analysis will score the post as enormously negative, and you can create alerts for posts with hyper-negative sentiment scores.

1.1.1Sentiment Analysis is hard

Today, Sentiment analysis plays an important role where various machine learning technique is used in determining the sentiment of very huge amounts of text or speech. Various application tasks include such as determining how someone is excited for an upcoming movie, correlates different views for a political party with people's positive attitude towards vote for that party, or by converting written hotel reviews into 5-star based on scaling across categories like 'quality of food', 'services', 'living room' and 'facilities' provided. As there is huge amount of information is shared on social media, forums, blogs, newspaper etc. it is easy to see why there is a need for sentiment analysis as there is much information to process manually which is not possible in today's time.

Main problem is that machine learning processes is not that much accurate. For a simple process to separate ‘positive’ view from ‘negative’ view on social media, many solutions can provide only performance around 80% accuracy. It can be useful to the track broad trends over time, but it may limit analysis of fined grained.

1.1.2Measuring Accuracy of Sentiment Analysis

The accuracy of Sentiment Analysis can be calculated in many ways, but the most common approach is to score accuracy in comparison to a human. So any natural language processing engine that scores around 80% is consider great job with accuracy. There are few major challenges for an engine analyzing text for sentiment. One of the biggest issues is that it has trouble understanding satire. Even humans have difficulty with someone who is being ironic. It is one of the most common mistakes a text analytics engine makes when trying to analyze text for sentiment.

Even humans have trouble, as they can analyze with 80% accuracy. Other problems are when words have multiple definitions. There are a few engines that use deep learning to help them understand context.

1.1.3Basic Component of an Opinion

It includes:

Opinion holder: Reviewer or an organization which holds a definite opinion on a particular or specific object or word. Object: it is considered a word, phrase or sentence on which an specified opinion is expressed. Opinion: refers an attitude, review or appraisal on particular object from the side of an opinion holder [23].

1.1.4Text Analysis process

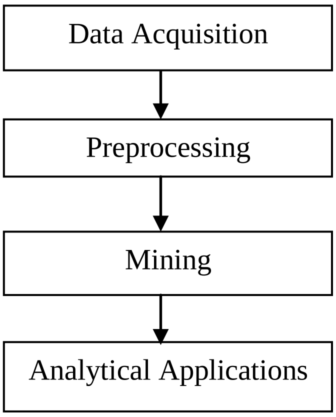

Fig 1.1: Process of analyzing text

This process involves the following steps [23]:

Data Acquisition: In this data acquisition, data are gathered from different relevant sources such as web crawling, twitter tweets, online review, newsfeeds, document scanning etc.

Preprocessing: It is used to remove noisy, inconsistent and incomplete data. For doing the classification, Text preprocessing and feature extraction is a preliminary phase.

Preprocessing involves 3 steps:

Tokenization or segmentation: It is the process of splitting a string of written language into its words. Text data consists of block of characters referred to as tokens. So the documents are being separated as tokens and have been used for further processing.

Removal of stop words: Stop words are the words which are needed to be filtered i.e. may be before or after natural language processing. Stop words are words which contain little informational. Various tools specifically avoid to remove these stop words in order to support phrase search. Several collections of words can be chosen as stop words for any purpose. Some search engines, removes most of the common words which include lexical words such as "want" from a text in order to improve performance. Search engine or natural language processing may contain a variety of stop words. It includes English stop words such as “and”, “the”, “a”, “it”, “you”, “may”, “that”, “I”, “an”, “of” etc. which are considered as ‘functional words’ as they don’t have meaning.

Researchers have shown that by removing stop words from the file, you can get the benefit of reduced index size without much affecting the accuracy of a user's. But care should be taken however to take into consideration the user's needs. Mostly, all search engines helps in eliminating the stop words from their indexes. With the help of eliminating stop words from the index, the index size can be reduced to about 33% for a word level index. While assessing the content of natural language processing, meaning of word can be conveyed more clearly by removing the functional word [21].

Stemming: It is the term which used to describe the process to reduce derived words to their origin word stem. Since 1960s, algorithms for stemming have been studied in the field of computer science. Different Stemming methods are commonly referred as stemming algorithms or stemmers. For English, the stemmer example are that, it should identify the string “cats”, ”catty” as based on the root word “cat”, and also “walks”, “walked”, “walking” as based on the root word "walk" [22].

Data Mining: Applying different mining techniques to derive usefulness about stored information. Different mining approaches are classification, clustering, statistical analysis, natural language processing etc. In text analytics, mainly classification technique is used. Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classified instance set. Data classifying and identifying is all about to tag the data so it can be create quickly and efficiently. But various organizations can gain from re-transforming their information, which helps in order to cut storage and backup costs, with increasing the speed of data searches. Classification can help an organization to meet authorized and regulatory requirements to retrieve specific information within a specific time period, and this is most important factor behind implementing various data classification technology.

Analytical Application: It provides valuable things from text mining so that it can provide information that helps in improving decision and processes. It includes following ways such as sentiment analysis, document imaging, fraud analysis etc.

2. PROBLEM

Different types of data are generated from different Social media groups that need to be organized and to monitor people's attitude towards products, gadgets, movie review etc. This database is collected from different social media sites for example Twitter, Face book, Online review, shopping sites etc. Text analytics and Sentiment analysis can help to develop valuable business insights from text based contents that may be in the form of word documents, tweets, comments and news that related to Social media. The foremost reason of Sentiment analysis is so complex is that words often take different meanings and are associated with different emotions depending on the domain in which they are being used. Dataset is analyzed by using the weka tool. The hidden relationship has to be extracted from this type of database using different mining approaches in Weka tool. Dictionary building for detailed sentiment analysis implies making an initial list of adjectives and nouns which are normally used when describing a specific movie review. Phrases and terms are extracted from this relational dataset and their meaning has been added to dictionary for next generation analysis. In tweets, informal and shortcuts has been used for explaining terms or views and this is done with the help of sentiments analysis is not an easy process. To reduce this, data mining approaches has been used for extraction of features from these datasets.

3. METHODOLOGY

In this proposed system, Weka an open source data mining tool has been used so as to perform sentiment classification on movie review dataset. Here, goal is to classify dataset into positive and negative and form the combined dictionary of Twitter dataset and online review dataset. Main steps are:

3.1Generating Dataset

Two dataset were collected firstly, from Twitter tweets and secondly, from Online review Dataset. The online review dataset consists of around 800 user's review archived on the IMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review were collected and each review were formatted according to .arff file where review text and class label are only two attributes. Class label represent the overall user opinion. Here, we set simple rules for scaling the user review. For dataset, a user rating greater than 6 is considered as positive, between 4 to 6 considered as neutral and less than 4 considered as negative.

3.2Preprocessing

For doing the classification, Text preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps:

I. Word parsing and tokenization: In this phase, each user review splits into words of any natural processing language. As movie review contains block of character which are referred to as token.

II. Removal of stop words: Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of preprocessing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

III. Stemming: It is defined as a process to reduce the derived words to their original word stem. For example, "talked", "talking", "talks" as based on the root word "talk". We have used Snowball stemmer to reduce the derived word to their origin.

3.3 Classification

Classification is a supervised learning method that helps in assigning a class label to an unclassified tuple according to an already classified instance set. Here, naïve bayes multinomial classifier has been used. Quality measure will be considered on the basis of percentage of correctly classified instances. For the validation phase, we use 10-fold cross validation method. Naïve bayes multinomial helps in generating dictionary and frequent set. It counts the occurrences of words in whole dataset and forms a dictionary of some most frequently occurring words.

Attributes are referred as text positions, values are referred as words.

$$c_{NB} = arg \max_{ci=C} P(c_j) \prod_i P(x_i|c_j)$$

From training the corpus huge amount of data, extract Vocabulary. Calculate the probability of $P(c_j)$ and $P(x_k \mid c_j)$ terms. For each c_j in C do. $docs_j$ which is referred as the total number of documents for which the target class is c_j.

$$P(c_j) \leftarrow \frac{| docs_j |}{| \text{total\# documents} |}$$

$Text_j$ it is referred as single document containing all $docs_j$ .For each word x_k in *Vocabulary*. n_k ← number of occurrences of x_k in $Text_j$.

$$P(x_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha \mid Vocabulary \mid}$$

positions ← all word positions in current document which contain tokens found in *Vocabulary*. Return c_{NB}.

3.4 Parameter Evaluation

Now for evaluating the result, different parameter are to be calculated. True positive, True negative, False positive and False negative are used for comparing the class label that have been assigned to a document by the classifier with the classes the item actually belongs [19].

Accuracy: It is measured as the proportion of correctly classified instances to the total number of instances being evaluated. Classification performance being evaluated by using this parameter [19].

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Where (TP)True positive – that are truly classified as positive.

False positive (FP)- not labeled by the classifier as positive but should be.

True negative (TN)- that are truly classified as negative.

False negative (FN)- not labeled by the classifier as negative but should be.

Precision: It is widely used in evaluating the performance in different field such as text mining, information retrieval. Precision is also referred to measure the exactness. It is defined as ratio of the number of correctly labeled as positive to the total number that has been classified as positive [19].

$$precision = \frac{TP}{TP + FP}$$

Recall: It is also used in evaluating the performance for text mining and information retrieval. It is also used to measure the completeness of the model. It is defined as the ratio of the number of correctly labeled as positive to the total number that are truly positive [19].

$$recall = \frac{TP}{TP \ + FN}$$

F-measure: It is referred as the harmonic mean of precision and recall. It helps to give score needed to balance between precision and recall. It combines two of them into one for the convenience as it might optimize the system so that it can favour one of them [19].

$$f = \frac{2 \times precision \times recall}{precision + recall}$$

3.5 Combined Dictionary

Combined word of twitter dataset and online review dataset forms a dictionary. As after classifying each word probability as positive, negative and neutral. Compare the probability for each word and categorize each word into three different dictionaries based on highest polarity of each word.

4. EXPERIMENTAL RESULTS

The online review dataset consists of around 800 user's review archived on the IMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review were collected and each review were formatted according to .arff file where review text and class label are only two attributes. Here, we analyse the dataset based on accuracy given by naïve bayes multinomial. Online review dataset accuracy around 94.968% and for twitter its around 82.695%. Results show that we get better accuracy for online review as compared to twitter tweets as online review are more clear and in detail compare to twitter tweets.

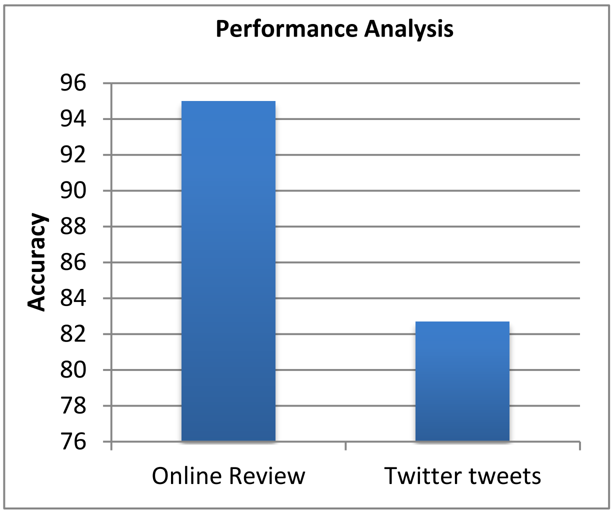

Fig4.1: Performance analysis based on accuracy

Combined dictionary of words of twitter tweets and online review are formed based on probability of each word as we get by classification algorithm i.e. naïve bayes multinomial.

5. CONCLUSION AND FUTURE SCOPE

Social media Monitoring has been growing very rapidly so there is a need for various organizations to analyze customer behavior or attitude of particular product or any movie review. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis on Online review are done by forming dictionary which shows that it is easier to build dictionary on phrases but complex in case of Twitter as tweets consist of short hands as online review were written in more clear way as compared to Tweets. So, form hidden relationship between different keywords and a dictionary of the words on the basis of different categories of comments & tweets.

Future work include to determine their features for the movie in detail i.e. make polarity check on different features such as actors, directors, scripts, music etc. and make the dictionary for them.

6. ACKNOWLEDGEMENT

I owe a great many thanks to a great many people who helped and supported me in writing this research paper.

7. REFFERENCES

[1] Ana Mihanovic, Hrvoje Gabelica, ZivkoKrstic 2014 Big Data and Sentiment Analysis using Knime: Online Reviews Vs. Social Media, MIPRO Opatija, Croatia.

[2] Basant Agarwal, Narmita Mittal, Erik Cambria. 2013 Enhancing Sentiment Classification Performance using Bi-tagged Phrases, 13th International Conference on Data Mining Workshops, IEEE.

[3] Bogdon Batrinca, Philip C. Treleaven 2014 Social media analytics: a survey of techniques, tools and platform Department of Computer Science, Gower Street, London, UK published in Springer.

[4] Deptii D.Chaudhri, R.A. Deshmukh. 2012 Feature – based Approach for Review mining Using Appraisal Words, Department of Post Graduate Computer Engineering, Pune, India, IEEE.

[5] Haruna Isah, Paul Trundle, Daniel Neagu. 2014 Social media Analysis for product Safety using Text mining and Sentiment Analysis, Artificial Intelligence Research Group, University of Bradford, UK, IEEE.

[6] Jalaj S. Modha, Gayatri S. Pandi, Sandip J. Modha. 2013 Automatic Sentiment Analysis for Unstructured Data, published in IJARCSSE.

[7] Javier Conejero, Peter Burnap, Omer Rana, Jeffery Morgan 2013 Scaling Archied Social Media Data Analysis Using a Hadoop Cloud sixth international conference on cloud computing, IEEE.

[8] Jiao Wu, Bin Zhang, Weihua Gao, Yi Hu, Jinsong Liu 2011,Online Web Sentiment Analysis on Campus Network, 4th International Symposium on Computational Intelligence and Design, IEEE.

[9] Kristin Glass and Richard Colbaugh 2012 Estimating the sentiment of social media content for security informatics applications, Institute for Complex Additive System Analysis, Socorro, USA, Springer Journal.

[10] Lingyan Ji, Hanxiao Shi, Mengli, Mengxia Cai, Peiqi Feng. 2010 Opinion Mining of Product reviews based on semantic role labeling, 5th International Conference on Computer Science and Education, IEEE.

[11] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt 2013 Big Data Privacy Issues in Public Social Media, Distributed Computing & Security Group, Leibniz Universitat Hannover, Thailand, Germany IEEE.

[12] Mohsen Farhadloo, Erik Rolland. 2013 Multi-class Sentiment analysis with clustering and score representation, 13th International Conference on Data mining Worshops, IEEE.

[13] Mrs. R.Nithya, Dr. D.Maheshwari. 2014 Sentiment Analysis on Unstructured Review, International Conference on Intelligent Computing Application, IEEE.

[14] Ms. K. Mouthami, Ms. K.Nirmala Devi, Dr. V.Murali Bhaskaran. 2010 Sentiment Analysis and Classification based on Textual Reviews, Dept of CSE, Tamil Nadu, IEEE.

[15] Nargiza Bekmamedova, Graeme Shanks 2013 Social Media Analytics and Business Value: A Theoretical Framework and Case Study, Department of Computing and Information Systems, University of Melbourne.

[16] Simona Vinerean, Iuliana Cetina 2013 The Effects of Social Media Marketing on Online Consumer Behavior, International Journal of Business and Management; Vol. 8, No. 14.

[17] SitaramAsur, Bernardo A.Huberma 2012 Predicting the Future with Social Media, Social Computing Lab, HP Labs, Palo Alto, California.

[18] V.K. Singh, R.Piryani, A. Uddin, P.Waila. 2013 Sentiment Analysis of Movie Reviews, Department of Computer Science, New Delhi, India, Published in IEEE.

[19] V. S. Jagtap, Karishma Pawar. 2013 Analysis of different approaches to Sentence-Level Sentiment Classification, published in IJSET.

[20] Ya-Ting Chang, Shih-Wei Sun 2013 A Real time Interactive Visualization System for Knowledge Transfer from Social Media in a Big Data, Center for Art and Technology, Taipei National University of the Arts, Taipei, Taiwan, IEEE.

[21] Stop words: http://en.wikipedia.org/wiki/Stop_words

[22] Stemming: http://en.wikipedia.org/wiki/Stemming

[23] Decideo Amérique Du Nord Inc www.decideo.fr/bruley/docs/2sentiment_a_v0.ppt