International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

A Study on the Techniques of Sentiment Analysis for Unstructured Data using Big Data Analytics

Renuka Devi D, Dr. Swetha Margaret T A
Asst. Professor, Department of Computer Science, Stella Maris College, Chennai, India

ABSTRACT

The real time unstructured data often refers to the information that doesn't follow the conventional storage of information in a row-column database. Unlike structured data it does not fit into relational databases. It is responsible for the Variety, one of the four V's of Big Data. Sources like satellite images, sensor readings, email messages, social media, web blogs, survey results, audio, videos etc., follow unstructured data. Organizations go beyond "basic" analytics and dive deeper into unstructured data to do things such as predictive analytics, temporal and geospatial visualization, sentiment, and much more. The objective of this paper is to confer model of sentiment analysis and its various techniques. Future research directions in this field are determined based on opportunities and several open issues in Big Data analytics.

Keywords: mining, Sentiment analysis, Unstructured data, Big Data

I. INTRODUCTION

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes, or emotions towards certain entities. An enormous growth of the WWW has been instrumental in spreading social networks. Due to many-fold increase in internet users taking to online reviews and opinions, the communication, sharing and collaboration through social networks have gained importance. The rapid growth in web-based activities has led to generation of huge amount of unstructured data which accounts for over 80% of the information. Exploiting big data alternatives in storing, processing, archiving and analyzing this data becomes increasingly necessary. In this paper we propose a generalized approach to analyzing sentiments in big-data environment.

II. LITERATURE REVIEW

One fundamental problem in sentiment analysis is categorization of sentiment polarity [1]. Given a piece of written text, the problem is to categorize the text into one specific sentiment polarity, positive or negative (or neutral). Based on the scope of the text, there are three levels of sentiment polarity categorization, namely the document level, the sentence level, and the entity and aspect level [2]. The document level concerns whether a document, as a whole, expresses negative or positive sentiment, while the sentence level deals with each sentence's sentiment categorization; The entity and aspect level then targets on what exactly people like or dislike from their opinions. Hu and Liu [3] summarized a list of positive words and a list of negative words, respectively, based on customer reviews. The positive list contains 2006 words and the negative list has 4783 words. Both lists also include some misspelled words that are frequently present in social media content. Sentiment categorization is essentially a

classification problem, where features that contain opinions or sentiment information should be identified before the classification. For feature selection, Pang and Lee [4] suggested to remove objective sentences by extracting subjective ones. They proposed a text-categorization technique that is able to identify subjective content using minimum cut. Gann et al. [5] selected 6,799 tokens based on Twitter data, where each token is assigned a sentiment score, namely TSI(Total Sentiment Index), featuring itself as a positive token or a negative token. Specifically, a TSI for a certain token is computed as:

$$TSI = \frac{p - tp / tn * n}{p + tp/tn * n}$$

where,

p is the number of times a token appears in positive tweets and n is the number of times a token appears in negative tweets. tp/tn is the ratio of total number of positive tweets over total number of negative tweets.

III. SENTIMENT ANALYSIS

People express opinions in complex ways, which makes understanding the subject of human opinions a difficult problem to solve. Rhetorical devices like sarcasm, irony, and implied meaning can mislead sentiment analysis, which is why concise and focused opinions like product, book, movie, and music reviews are easier to analyze. From a user's perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites.

For a case study, a new mobile phone launched in the market and the company would like know what feature of your phone people are liking or disliking. Consider the following comment "The phone is just awesome. I have this phone using from 2 weeks, it is the best phone and great value for money Brilliant Display-8.5/10. Operation is fast-8/10(A few lags after using it for a month). Good Gaming Experience-8/10. Good Battery Backup-8/10(1day 15hrs on medium usage). Looks Elegant-8/10. Value for Money-9.5/10. Heats up While Playing Games for more than 30 minutes. Average Camera Performance-6/10(flash is extremely powerful though). Limited Onboard Storage (Enough for me though). Overall-9/10(Killer Price--Not a Single phone within 20k can match to its performance)"

In the above sentence user has commented almost on all the major attributes of a Mobile.

The approach to find sentiment analysis is basically following:

- Identify the entity (in this case it is a Mobile Phone).
- Identify attributes of the entity (in this case they are : Display, Performance, Battery Backup, Looks, Camera, Storage, Price)
- Find Sentiment corresponding to each attribute (in this case user has a positive sentiment/review for display, performance, backup, look and somewhat negative sentiment or neutral for Camera)

IV. METHODOLOGY FOR SENTIMENT ANALYSIS PROCESS ON PRODUCT REVIEWS

The proposed model is shown in the figure 1.

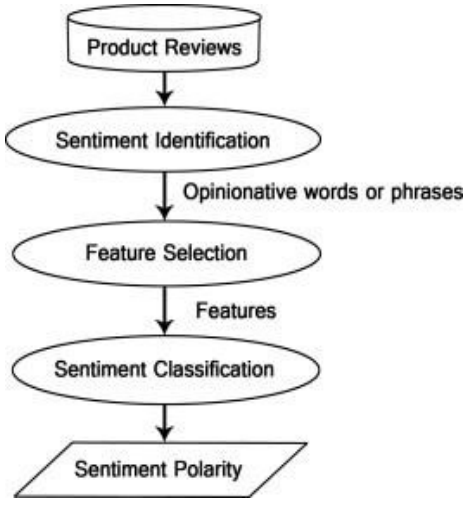

Figure 1. Model

4.1. Sentiment identification

In the case of small amount of text it is possible to read through it and look for sentiment-bearing phrases i.e. the noun phrases within the text. Then personal judgement is used to decipher the sentiment of the phrase and create a dictionary of all sentiment bearing phrases. Instead of this traditional methodology of discernment, we can use a sentiment analysis engine to analyze a large amount of text. Natural Language Processing by Steven Bird, Ewan Klein, and Edward Loper [6] is the definitive guide for NLTK, walking users through tasks like classification, information extraction. It is an open source tool for taming texts. This tool works by examining individual words and short sequences of words (n-grams) and comparing them with a Bayesian probability model [7]. It can detect negations in phrases, i.e, the phrase "not bad" will be classified as positive despite having two individual words with a negative sentiment.

4.2. Feature Selection

Sentiment Analysis task is considered a sentiment classification problem. The first step in the Sentiment classification problem is to extract and select text features. Some of the current features are:

Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features as discussed earlier.

Parts of speech (POS): Finding adjectives, as they are important indicators of opinions.

Opinion words and phrases: these are words commonly used to express opinions including good or bad, like or hate. On the other hand, some phrases express opinions without using opinion words.

Negations: The appearance of negative words may change the opinion orientation like not good is equivalent to bad.

4.2.1. Feature selection methods

Lexicon based methods, statistical methods and machine learning techniques are the most frequent Feature Selection methods.

4.3. SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach [8].

4.3.1. Machine Learning: Machine Learning is to automatically learn to make predictions on current data based on past history. In machine Learning you need to have a training set which is nothing but already classified comments and you use any classification algorithm like naive Bayes, SVM etc.,[10] to train a classifier using the training set. And finally you can use the classifier to find sentiments of new comments.

4.3.2. Lexicon based approach: In lexicon analysis you need to maintain list of positive and negative sentiment words. You read comments line by line and search for sentiments words and maintain a running score for positive and negative words, if positive score is greater it is classified as positive comments and vice-versa. Sentiment Analysis on feedback comments are bit tricky, usually the user who's writing his comments do not write grammatical or lexical correct English along with various short language. For example, Sarcasms are so common in consumer reviews about products and services, which make opinions hard to deal with.

Words like good, wonderful, and amazing are positive sentiment words, and bad, poor, and terrible are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., cost someone an arm and a leg. A list of such words and phrases is called a sentiment lexicon (or opinion

lexicon). Over the years, researchers have designed numerous algorithms to compile such lexicons. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis. Below, we highlight several issues:

A positive or negative sentiment word may have opposite orientations in different application domains. For example, "suck" usually indicates negative sentiment, e.g., "This camera sucks," but it can also imply positive sentiment, e.g., "This vacuum cleaner really sucks."

A sentence containing sentiment words may not express any sentiment. This phenomenon happens frequently in several types of sentences. Interrogative sentences and conditional sentences are two important types, e.g., "Can you tell me which Sony camera is good?" and "If I can find a good camera in the shop, I will buy it." Both these sentences contain the sentiment word "good", but neither expresses a positive or negative opinion on any specific camera. Sarcastic sentences with or without sentiment words are hard to deal with, e.g., "What a great car! It stopped working in two days."

4.4. SENTIMENT POLARITY

It means emotions expressed in a sentence.

Emotions are closely related to sentiments. The strength of a sentiment or opinion is typically linked to the intensity of certain emotions, e.g., joy and anger. Opinions in sentiment analysis are mostly evaluations. According to consumer behavior research [9], evaluations can be broadly categorized into two types:

- ✓ Rational evaluations
- ✓ Emotional evaluations.

Rational evaluation: Such evaluations are from rational reasoning, tangible beliefs, and utilitarian attitudes. For example, the following sentences

Express rational evaluations: "The voice of this phone is clear," "This car is worth the price," and "I am happy with this car."

Emotional evaluation: Such evaluations are from non-tangible and emotional responses to entities which go deep into people's state of mind.

For example, "I love iPhone," "I am so angry with their service people" and "This is the best car ever built."

To make use of these two types of evaluations in practice, we can design 5 sentiment ratings, emotional negative (-2), rational negative (-1), neutral (0), rational positive (+1), and emotional positive (+2). In practice, neutral often means no opinion or sentiment expressed.

V. CONCLUSION AND FUTURE WORK

Recognizing semantic and linguistic features instead of just the statistical significance of words can enhance performance of the system. Also, the iterative scaling can be improved by using Improved Iterative Scaling [Berger, 1997]. Since the corpora is small, the classification accuracy can be somewhat improved by grouping users with similar tastes

VI. REFERENCES

[1]. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends 2(1-2):1–135

[2]. Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers

[3]. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA. pp 168–177

[4]. Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In:

Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04. Association for Computational Linguistics, Stroudsburg, PA, USA

[5]. Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system. In: Proceedings of the second ASE international conference on Big Data. ASE

[6]. Analyzing Text with the Natural Language Toolkit by Steven Bird, Ewan Klein, and Edward Loper.

[7]. Naive_Bayes and Sentiment classification. In Stanford University.

[8]. Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: Proceedings of the 8th international conference on the semantic web, ESWC'11; 2011. p. 88–99.

[9]. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2014.

[10]. Comparison of Classification Algorithms in Text Mining. International Journal of Pure and Applied Mathematics [Volume 116 No.22 2017, 425-433]