International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 6, Issue 11
November 2017

# Mining Social Media Data to Identify the Sentiments of Reviews

**Subhash Bhagavan Kommina**
Assistant Professor,
Sasi Institute of Technology & Engineering, Andhra Pradesh, India

**P V P Sandeep**
Sasi Institute of Technology & Engineering, Andhra Pradesh, India

**V Salman Raj**
Sasi Institute of Technology & Engineering, Andhra Pradesh, India

***Abstract:*** *Internet is a collection of rich source of information where users share their opinion of a particular product, general activities global issues and others. By analyzing the text posted by users in web, it will be helpful for a service sector and business to know the opinion of the customer of their products from the feedback. Sentiment analysis is an important research area in text mining, in which the sentiments are gathered, analyzed and grouped from the text. It plays a prominent role in academics and business domain. In this paper, sentiments are identified using a classifier named as Logistic Regression combined with Natural Language Processing Techniques to improve the accuracy of the classifier. This model worked on Amazon review dataset and identified the polarities of the products using the classifier.*

***Keywords: Sentiment Analysis, Natural Language Processing, Logistic Regression, Text Mining.***

## 1. Introduction

The electronic Word of Mouth (eWOM) statements expressed by the users of the products help the organizations to analyze and to identify the opinion of the user of their products. Most of the people share their opinion about the products in social media like twitter, Face book and others. . To analyze such huge data we need some automated Text Mining techniques to predict the polarities of text. To do so Sentiment analysis is one of the branch of Text Mining this is used to find the sentiments of the text, which is used for business development. Social media provides a better means of expressing people's views and ideas on different issues like politics, trends of society, events, academics and others.

Sentiment analysis is also known as opinion mining[1]. This is the technique used to resolve the problems encountered by the users with huge number of reviews. Here we will analyze the text and say whether the given review is either positive or negative.

There is a huge amount of data in the form of reviews on the web which talks about user feedback [2]. It is very hard to generalize and summarizing the opinions of the text in social media posted by different diversity of people. Most of the organizations require automated text analyzer or miner to do this process. Every business organization is adopting sentiment analysis tools and techniques to analyze the text posted by their customers [3].

## 2. Literature review

The objective of this survey is to project the important research studies in the area of sentiment analysis. The literature is concerned on sentiment analysis tasks

### 2.1. Subjectivity classification

Subjective classification is used to classify the text either subjective or objective. These kinds of classifications can be done using Machine learning techniques.

In this aspect, Abdul-Mageed et al. [7] proposed a machine learning approach to classify subjectivity and sentiment at sentence level . His work firstly on newswire documents from PATB and the authors collected and annotated around 11,918 sentences from different social media services. Here classification is effected in two stages. In the first , subjectivity classification such as a distinction is drawn between a subjective and objective text. In the second , sentiment classification is done by differentiating positive and negative sentiment. Here they proposed SVM machine learning algorithm along with language specific and general features. Features like n-grams , domain polarity

and unique lexicon features. He also added POS tagging and lemmas to extract the positive words that are impact on subjectivity and sentiment classification.

Nabil et al. [8] presented a 4 way classification of sentiment using 4 classes i.e., Objective, subjective negative, subjective positive and subjective mixed. Their data set contains manually annotated 10,006 tweets using Amazon Mechanical Turk(AMT) service. The algorithms applied by them are SVM , MBN, BNB , KNN on both balanced and unbalanced datasets. Even using n-grams in multi-way classification they didn't give accurate results

A lexicon-based approach was proposed in [9] to perform the subjectivity classification of both MSA news articles and micro blogs from twitter. To build a large lexicon , the authors will use two lexicons : MPQA (English subjectivity lexicon) and ArSenti an Arabic lexicon. The first is translate into Arabic using Machine Translations and the second is extension of random walk graph method. Here every word is to be tokenized and stemmed.

### 2.2. Sentiment classification

This classification's main aim is to classify the subjective texts in more than two types. The first is binary classification contains only positive or negative opinion. A multi way classification labels texts according to strength of the sentiment in that word like extremely negative , negative , neutral , positive , extremely positive.

Mountassir et al. [10] conducted one binary sentiment classification using NB , SVM , KNN. Here they used two corpora's: The first contains two domain specific datasets .The second is OCA on movie reviews developed by Rushdi-saleh et al.

[11]. Before the classifying here authors performed a preprocessing task by removing stop words , eliminating terms only once or twice in the data set and by replacing with their stems. Here usage of n-grams and weighting improve the classification performance.

In [12], sentiment is classified in two forms :Polarity classification and the rating classification. Here the review is to be scaled from 1-5.

Aly and Atiya [12] created LABR a dataset over 63,257 book reviews collected from "goodreads". Reviews with 4 or 5 will be positive and 1 or will be negative and 4 will be neutral .They applied Machine Learning in both balanced and unbalanced data using SVM , MNB , BNB as algorithms and n-grams as features and achieved a good result up to 90% accurate.

El-Baltagy et al. [13] proposed a lexicon based approach to establish a sentiment classification. After building a lexicon of 4392 terms , the authors used two data sets. The first calculates one score for each document by adding weights of negative and positive terms. The second is assigning a positive and negative to each lexicon. The authors achieved two algorithms on a Twitter dataset up to 83.8% accurate.

In [14], El-Makky et al. Combined Sentiment Orientation algorithms with a machine learning classifier to build a hybrid approach. They used Twitter data set used lexicon based approach. The scores were integrated with different polarity features with 84%accuracy.

## 3. Related Work

Sentiment analysis is performed at three levels, they are document level, sentence level and aspect level. At document level sentiment analysis the whole document is considered as a single unit and the document will be classified as with polarities such as positive, negative and neutral.The other level is sentence level in which every sentence will be considered as a part and identifies either the sentence is holding an opinion or not. Sentence level sentiment analysis is more challenging when considered the level of granularity of words with highly context dependent. Aspect level sentiment Analysis performs fine grained analysis of text based on a *"quintuple (o; a; s0; h; t)"* of the opinion. There are mainly two important steps they are aspect extraction and aspect sentiment classification.

### 3.1. Steps to do Sentiment Analysis

#### 3.1.1. Data Collection using Web Scraping

In order to perform real time sentiment analysis, first step is to collect data from web. It includes the technique called web scrapping used to collect reviews from websites. As the data is in online we have to get the data from the online and this process is called Data Scrapping. The scrapped data may be a text or multimedia like audio or video etc. In this paper we will discuss only about text mining but not about multimedia.

For example twitter is providing Twitter Rest API to get the data about the users and profile

Academic Science

International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 6, Issue 11
November 2017

information etc. These API 's will help developers to get the required data from the web [10].

### 3.1.2. Preprocessing

The data sets those are available after scrapping is to be go through some of the preprocessing techniques before we actually analyze the data. Some of them are using Ngrams, tokenizing, POS tagging, Stemming, Lemmatizing, SentiWordNet and others[5].

#### 3.1.2.1. Using Ngrams

Here the review or the text we have is to be tokenized into single words known as 'Unigrams'. But in case of unigrams we have some of the problems like follows. Let us consider an example and we will know why we are going with these bigrams instead of unigrams."Food is good and service is not good." In this case number of positive words are 2 i.e., good in food aspect ,and good in service aspect and number of negative words are 1 i.e., not in service aspect. But actually the no of positive words are only 1 and not god is to be considered as a single entity i.e., as negative word. So here when we are changing our text tokenizing from unigram to Bigram we will tokenize every two words as a single entity. So that not good is considered as a single entity .

This unigrams , bigrams are extended up to Ngrams , so that the tokenizing of grams increases we can accurately give the results.

#### 3.1.2.2. Tokenizing

Splitting of the text we have is known as tokenizing. This tokenizing may be done unigram or bigram or trigram and so on.

#### 3.1.2.3. POS Tagging

The text we have is to be assigned with Parts Of Speech and this is useful in our analyzing part as, in general the aspect we will consider is a noun and the polarity is to be defined as an adjective. So in order to do ABSA we have to know the POS of every word in the text. So we will tag the POS to every word here [1][2].

#### 3.1.2.4. Stemming and Lemmatizing

Stemming and Lemmatizing is used to reduce the forms of words into their root word. It changes any form of the word to root form.

Example: car , cars , car's , cars' => car.

#### 3.1.2.5. SentiWordNet

SentiWordNet is most powerful lexical resource used in Sentiment Analysis. It is a sysnset of Word Net which majorly allocates three classes to the sentences given as Postive, Negative and Neutral [5][6].

For example: 'The' = neutral , 'good' = positive , 'bad' = negative.

### 3.2 Proposed Model

#### 3.2.1. Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics.

It is the go-to method for binary classification problems

#### 3.2.2 Logistic Function

The function used at the core of the method is called Logistic Function.

The "Logistic function" also known the "Sigmoid function". It's an S-shaped curve that can take any real-valued numeric and map it into a value 0 and 1 , but never exactly at those limits[20].

$$1 / (1 + e \wedge - v \quad )$$

Where e is given as a base for nlp algorithms and value obtained is a number which is transformed from a text.

#### 3.2.3 Logistic Regression Classifier (LRC)

LRC uses logistic equation to compute the values and this model is similar to linear regression technique.

Input: It takes values with weights (x)

Output: It predicts the value (y).

The major difference between LRC and linear regression is the output value is modeled as a binary value rather than a numeric value.

This equation refers to Logistic Regression where y is a predicted output, where b0can be a bias and b1 will be a coefficient value for the input term x.

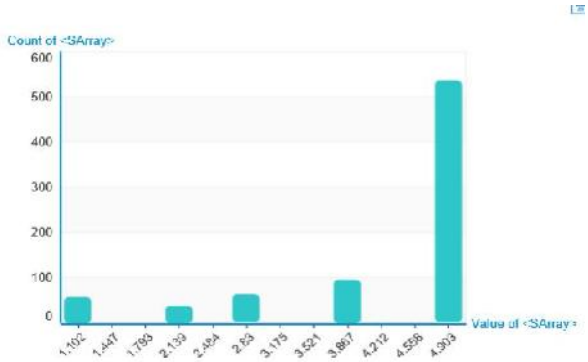
$$y = e \wedge (b0 + b1 * x) / (1 + e \wedge (b0 + b1 * x))$$

## 4. Results

After constructing the above said matrices we apply Logistic Regression Classifier (LRC) at this stage to evaluate the performance of the classifier. To evaluate the model Amazon review data set is considered which consist of reviews of different items and their review ratings.

International Journal of Innovations & Advancement in Computer Science
IJIACS
ISSN 2347 – 8616
Volume 6, Issue 11
November 2017

fig 5.1 shows the distribution of the reviews and represented using sparse array .




To measure the accuracy of a model for predicting the sentiment of the review , the following True-Positive(TP) , True-Negative(TN) , False-Positive(FP) and False-Negative(FN) are taken into consideration for constructing confusion matrix. In Sentiment Analysis the factors to e considered are Precision, F-score, Recall are used for evaluating the model for predicting the polarity of the text [2].

Confusion matrix is used to estimate the accuracy of the classifier. It will be matrix representation of showing the actual values and predicted values of a classifier.

| | | **Actual Class** | |
|---|---|---|---|
| **Predicted Class** | | Actual :**TRUE** | Actual :**FALSE** |
| | Predicted: **POSITIVE** | 26521 | 1327 |
| | Predicted: **NEGATIVE** | 4001 | 1455 |

Confusion matrix for Amazon reviews

The Precision (P) Measure is given as

$$P \quad (P) = T \ / (T \ + F \ )$$

The Recall (R) Measure is given as

$$R \quad (R) = T \ / (T \ + F \ )$$

The F-score (F) Measure is given as

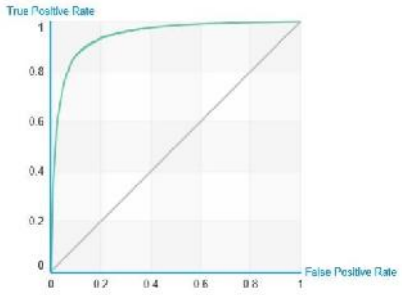
$$F - s \quad (F) = (2 * P * R) / (P + R)$$

The below table shows the evaluation measures for the model and the results

Table 5.1 Model evaluation parameters

| Evaluation parameter | Value obtained in % |
|---|---|
| ACCURACY | 91.6 |
| PRECISION | 95.2 |
| F- SCORE | 95 |
| RECALL | 94.8 |

AUC Curve is termed as Area Under Curve which is used to drawn a random positive before a random negative. It show the true postive and false negative path under a threshold of 0.5




## 5. Conclusion

The main aim of this paper is to extract sentiment from the given text using LRC. In this paper the model is built by using both machine learning algorithms and Natural Language Processing (NLP). The model uses Logistic Regression in order to calculate the sentiment here. This paper it is concluded that using machine learning algorithms are efficient in feature extraction. Using this model we get an accuracy of 91.6% in analyzing the given text.

## 6. References

1. B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.
2. Kumar Ravi, Vadlamani Ravi , A survey on opinion mining and sentiment analysis: tasks, approaches and applications.
3. I. Niles, A. Pease, Linking Lexicons and ontologies: mapping Word Net to the Suggested Upper Merged Ontology, Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03), Las Vegas, 2003, pp. 23–26.
4. C. Strapparava, A. Valitutti, WordNet-Affect: an affective extension of WordNet, Proceedings of LREC, vol. 4, 2004, pp. 1083–1086.
5. A. Esuli, F. Sebastiani, SENTIWORDNET: a publicly available lexical resource for opinion mining, Proceedings of the 5th Conference on Language Resources and Evaluation LREC-06, Genoa, Italy, 2006, pp. 417–422.
6. S. Baccianella, A. Esuli, F. Sebastiani, SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion
7. SAMAR: Subjectivity and sentiment analysis for Arabic social media , Abdul Mageed
8. Arabic sentiment Tweets data set , Nabil et al.

9. A. Moreo, M. Romero, J.L. Castro, J.M. Zurita, Lexicon-based Comments-oriented News Sentiment Analyzer system, Expert Systems with Applications 39 (2012) 9166–9180.
10. C. Bosco, V. Patti, A. Bolioli, Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT, Knowledge-Based.
11. S.-T. Li, F.-C. Tsai, A fuzzy conceptualization model for text mining with application in opinion polarity classification, Knowledge-Based Systems 39 (2013) 23–33.
12. E. Boiy, M.-F. Moens, A machine learning approach to sentiment analysis in multilingual Web texts, Inf. Retrieval (2009) 12:526–558, DOI 10.1007/s10791-008-9070-z.
13. Y. Dang, Y. Zhang, H. Chen, A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews,Sentiment Classification, IEEE Intelligent Systems, July/August 2010.
14. T. Mullen, N. Collier, Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp.412-418), 2004, July.
15. A. Balahur, Methods and Resources for Sentiment Analysis in Multilingual documents of Different Text Types, PhD Thesis, University of Alicante, Spain, 2011, 273 pages.
16. Deep Learning for Aspect-Based Sentiment Analysis , Bo wang , Min Liu Department of Electrical Engineering , Stanford University.
17. Survey on Aspect-Level Sentiment Analysis , Kim Schouten and flavius Frasincar , IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 3, March 2016
18. Opinion Mining: Aspect Level Sentiment Analysis using SentiWordNet and Amazon Web Services , Kajal Sawargi , International Journal of Computer Applications (0975 – 8887) Volume 158 – No 6, January 2017
19. Classification Techniques for Sentiment Discovery-A Review , Salina Adinarayana , International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016
20. Maas, Andrew L., et al. "Learning word vectors for sentiment analysis." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.