A Peculiar Sentiment Analysis Advancement in Big Data

Manisha Valera¹, Yash Patel²

¹ Assistant Professor, Department of Computer Engineering, Indus University, Ahmedabad, Gujarat, India
² Student, Department of Computer Engineering, Indus University, Ahmedabad, Gujarat, India

Abstract. Every company wants to discover what its customers feel about it. But sentiment analysis can get coarser and turn inward to improve employee as well as customers’ satisfaction. A term called sentiment analysis, or the mathematical taxonomy of statements’ negative or positive connotations, gives companies potent ways to analyse cumulative language data across all sorts of communications. There's real value in measuring sentiment inside and outside your company. Our big data experts analyse the core grounds of the problems raised while handling giant datasets and find solutions to efficiently manage massive datasets with ease. We built a sentiment analysis solution to measure positive and negative sentiments with top rating products for Amazon.com’s Electronics products as against competitor products.

1. Introduction

Sentiment Analysis involves recognizing the evaluative nature of a part of passage or text. For example, a product review can express a positive, negative, or unbiased sentiment. Automatically discovering sentiment articulated in text has a number of applications, including tracking sentiment towards products, movies, diseases etc., improving customer relation models, detecting pleasure and well-being. Over the past decade, there has been an extensive escalation in the use of micro blogging services such as Twitter and access to mobile phones world-wide. Thus, there is incredible curiosity in sentiment analysis of short informal texts, such as tweets and SMS messages, across a variety of domains. Now a days Internet is ruling over each and every aspect of the world, either selling or buying, ticket booking, ecommerce, education, medical field. Today, if a customer wants to buy a single product, He or she goes to ecommerce sites or application on mobile and check the Product’s rating and review overall the world and determine which Product is bestselling in the compare of other brands. And then determine which one to buy. So, the review of the Product plays a vital role for getting sold the products. In our proposed algorithm we have taken Amazon data set of the Electronics Products and on that we have applied an algorithm to find out top rating products and on that we have applied sentiment analysis to find out the high positive sentiments from the customer reviews. So, the consequence of our algorithm will be highly selling Products all over the world in given data set which can be used for business analysis.

With the dawn of Big Data Analytics and Machine learning, it has been possible to create intelligent systems, which can analyse texts and derive insight which could be advantageous for business. In Ecommerce platforms with visitors hitting up to 2 billion, there is a lot of raw data that
can be analysed to understand the end user behaviour. If the user clicks, reviews, tags, blog posts, ratings all are gathered and processed many form of tortuous indicators can be converted to key performance indicators of a business thereby helping businesses to find target areas to be beneficial from manager perception. The customers are the solitary intention of any business; measuring their contentment level through these reviews can help both the customers and managers. Visuals for long have been making a lifelong impact on individuals, thereby empowering them in coming up with quick decisions. Providing them with visualizations and recommendations on the foundation of what others are saying can be relatively fruitful.

2. Related Work

In [1], presented a novel approach named Feature-Grading used to make the proposal of products in the e-commerce business. The procedure of Feature-Grading is isolated into five key strides: (a) Extracting general list of capabilities of a gathering classification of wares; (b) Extracting modifier set and negative words set; (c) Acquiring particular list of capabilities and highlight evaluation set; (d) Acquiring particular component weight set; (e) Acquiring thing weight set. The author has recommended the top ranking items using by applying rank and grade to all the items with an acquired grading equation.

A Proposed method in [2] is used to calculate sentiment score of reviews given by the customers is implemented in python on Hadoop. It uses Hadoop Distributed File System (HDFS) to store data set and run on Map Reduce architecture for performing sentiment analysis. The method works in two phases: In the first phase, Mapper will parse the given input file and in the second phase, Reducer will compute the sentiments. The data set they used contains different types of clothing product reviews from Amazon.com. Which present a statically a number of reviews about the product, number of positive and negative words, number of stop words. Stop word dictionary is used to recognize and remove stop words from the reviewed product. Increase/decrease the polarity strength when a word is preceded by a negation. Polarity Calculator uses positive, negative and stop-word dictionary to extract the sentiment-oriented words from each sentence. After calculating polarity the output will be stored in HDFS. Accuracy is performed by using the review scores of a product given by customers. The error rate for positive and negative reviews are comparatively little.

Shoushan [3] has performed a sentiment classification with polarity shifting phenomenon. They have extracted first the polarity shifted sentences using some detection rules. The same rules they have applied on testing data. And counting based classifier was designed based on polarity shifted words.

The author has proposed a Linguistic approach in [4], in that they have used emoticon symbols for separating the positive and negative comments in facebook. Thus, they have done Sentiment Lexicon analysis using sentiment words and classified positive and negative comments by prepared dictionaries.

The tweets were analysed in [5] to classify data and sentiments from Twitter more accurately. The information was extracted based on knowledge extraction. This is further enhanced using sphere specific seed based enhancement technique. The proposed method provides the extraction of keywords, entities, synonyms, and parts of speech from tweets which are then used for tweets classification and sentiment analysis. The proposed framework is tried on a gathering of 40,000 tweets. By applying the Knowledge Enhancer and Synonym Binder module on the extracted information the authors have achieved increase in information gain in a range of 0.1% to 55%.

In [6], it is intended to analyse social media big data to identify the widespread of certain keywords. A Java-Hadoop application is implemented to analyse data obtained from Twitter social network. The purpose was to identify number of people (Tweets) who mentioned specific medical keywords (e.g. Cancer) classified by location. The application is developed using Java Eclipse on CentOS Linux operating System and runs on Oracle Virtual Machine. It is aimed to help in decision making according to the number of people tweeting about cancer or any related word (like tumor) and analyse them according to their cities all over the world. The results are used to create a GIS layer to spatially boost the visualization of the obtained consequences.
The sentiment analysis system [7] that detects the sentiment of short informal textual messages such as tweets and SMS (message-level task) and the sentiment of a word or a phrase within a message. The system is based on a supervised statistical text classification approach leveraging a variety of surface form, semantic, and sentiment features. The sentiment features are mainly derived from tweet-specific sentiment lexicons. These lexicons are generated from tweets automatically with sentiment-word hash tags and from tweets with emoticons. To efficiently capture the sentiment of words in negated contexts, a separate sentiment lexicon is generated for negated words.

3. Proposed Work
In this section, a method to calculate the rating and based on that sentiment score of reviews given by the customers is proposed and implemented in python. This method works in two phases: In first phase, it will parse the given input file and in second phase, it will discover the average of each Product’s rating and based on that it will find out the highest rating as outcome and calculate the sentiment score of reviews given by the Customer. The data set we used contains different types of Electronic products’ rating and reviews from Amazon.com [8]. Some statistics about the data set are given in Table III. For instance, number of Products, number of Products with 5.0 rating, number of positive and negative sentiments. We use a positive and negative word dictionary to identify positive and negative words, Accuracy and Error rate [9]. Stop word dictionary is used to identify and remove stop words from the reviewed product [10].

4. Implementation and Result Analysis
The proposed approach is implemented on Amazon product reviews dataset [11]. We have implemented in python 3.5, the dataset which we have used is in JSON file format.

```json
{
  "reviewerID": "A2SFAM1J3QNN36",
  "asin": "000013714",
  "reviewerName": "J. McDonald",
  "helpful": [0, 1],
  "reviewText": "I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase thought!
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1528260000,
  "reviewTime": "59 13, 2018"
}
```

Figure 1. Sample Review

We have converted JSON file format to CSV file format by removing unwanted columns and fetched only required columns like reviewer ID, Product ID, reviewer Text, Overall rating using Panda framework for increasing loading and reading speed. So, the file size would be reduced by 21.8%.
After converting to CSV format,

- Total Number of Reviews available: 1626187
- Total Number of Products available: 63000

We read a CSV file’s Product ID, Overall rating into pandas data frame.

- Product wise group the data and find the average Overall rating of individual Product.
- Fetch the List of Product ID with Overall rating is 5.0.

After finding the Overall rating=5.0, we got total 2191 Product IDs. We have fetched reviewer Text for the products whose Overall rating=5.0.

- Analyze reviewer Text Product ID wise.
- Remove Stop Words.
Stop Words like which, after, has, that, become, in, before. are those words which are mostly used in a sentence and not of use in the sentiment analysis, so exclusion of these type of words will not affect the experimental results.

- Divide the reviewer Text into Positive and Negative Sentiments accordingly.

Usually negation will change the polarity of positive and negative words. For instance, “this product is not good” – in this sentence good is a positive word but if it is used with not it becomes negative[11]. So, we have given Weight age to each sentiment.

Table 1. Type of Sentiments

| Positive Sentiments          | Negative Sentiments          |
|-----------------------------|-----------------------------|
| 'good', 'nice',             | 'bad', 'not good',          |
| 'excellent', 'superb',      | 'worst', 'cheap',           |
| 'awesome', 'expensive',     | 'inferior'                  |
| 'fine', 'outstanding',      |                             |
| 'magnificent', 'fantastic', |                             |
| 'brilliant', 'robust', 'not bad' |                     |

Table 2. Weightage for Sentiments

| Sentiments                              | Weightage |
|-----------------------------------------|-----------|
| nice, good, fine, robust                | +1        |
| excellent                               | +2        |
| superb, fantastic, outstanding, magnificent, brilliant | +1.5   |
| not bad                                 | +0.9      |
| bad, cheap, inferior                    | -1        |
| not good                                | -0.9      |
| worst                                   | -2        |

- Sort the List of Product ID according to Weightage of sentiments.

Using sorting we got the Products with Highest Positive Reviews which have Overall rating=5.
Though Our Algorithm’s outcome has majority of Positive values for the sentiments, there is some human error for the sentiments which are having top rating for the products. Therefore, some products’ review shows negative value in our prediction.

Figure 4. Product IDs with Overall Rating is 5.0

| Figure 5. Sortest List of Product ID with Highest Review Weightage and Overall rating=5.0 | |
|---|---|

C:\Users\Gan\AppData\Local\Programs\Python\Python38\python.exe B:\Amazon\amazon.py
Table 3. Results after Implementation.

| Number of products | 63000 |
|--------------------|-------|
| Number of products with 5.0 rating | 2191 |
| Number of products with positive comments | 2085 |
| Number of products with negative comments | 106 |
| Accuracy (%) | 95.16 |
| Error (%) | 4.84 |

Figure 6. Number of Products with Overall rating=5.0

After Implementation of this algorithm, A test of the sentiment analysis has been conducted to quantify accuracy. There would be 95.16% accuracy in finding out which Product is best sold based on rating and review. But, due to human error in reviews we got error of 4.84%.

Figure 7. Number of Products with Overall rating=5.0 and Highest Positive Sentiments

5. Conclusion
Using Proposed Algorithm, We can find out the best selling product based on the rating as well as by sentiment analysis on reviews given by user, since first of all we have extracted the products with top rating, that is 5 star rated and after that by applying sentiment analysis on top fetched product for finding the most best reviewed products of all top rated product and sorting the products according to their Weightage of sentiment types, we got 95.16% accuracy. That means we are 95.16% successful in understanding human sentiment regarding top products and hence predicted the best product of all product. For further advancement in finding best product we can use machine learning algorithm and Natural Language processing for detecting weightage for sentiment keywords.
6. References

[1] Luo Yi; Fan Miao; Zhou Xiaoxia, *The design and implementation of Feature-Grading recommendation system for e-commerce*, Information and Automation (ICIA), 2011 IEEE International Conference on, vol., no., pp.236, 241, 6-8 June 2011

[2] Piyush Gupta, Pardeep Kumar, and Girdhar Gopal. *Article: Sentiment Analysis on Hadoop with Hadoop Streaming*. International Journal of Computer Applications 121(11), 4-8, July 2015

[3] Shoushan Li; Zhongqing Wang; Lee, S.Y.M.; Chu-Ren Huang, *Sentiment Classification with Polarity shifting Detection*, Asian Language Processing (IALP), 2013 International Conference on, vol., no., pp.129,132, 17-19 Aug. 2013

[4] Jamoussi, S.; Ameur, H., *Dynamic construction of dictionaries for sentiment classification*, Cloud and Green Computing (CGC), 2013 Third International Conference on, vol., no., pp.418,425, Sept. 30 2013-Oct. 2 2013

[5] Batool, R.; Khattak, A.M.; Maqbool, J.; Sungyoung Lee, *Precise tweet classification and sentiment analysis*, Computer and Information Science (ICIS), 2013 IEEE/ACIS 12th International Conference on, vol., no., pp.461,466, 16-20 June 2013

[6] Alenezi, Sultan, and Saleh Mesbah, *Big Data Spatial Analytics in Social Networks using Hadoop*, International Journal of Computer Applications 128.14 (2015): 21-26.

[7] Svetlana Kiritchenko, Xiaodan Zhu, Saif M. Mohammad, *Sentiment Analysis of Short Informal Texts*, Journal of Artificial Intelligence Research 50 (2014), 723–762

[8] jmcauley.ucsd.edu/data/Amazon/

[9] Minqing Hu and Bing Liu., *Mining and Summarizing Customer Reviews*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

[10] “stop-words,” (Last visited in June 2015) [online]. Available: https://code.google.com/p/stop-words/

[11] Godbole, N., Srinivasiah, M., & Skiena, S., *Large-Scale Sentiment Analysis for News and Blogs*. ICWSM 7 (2007): 21