Sentiment Analysis of Customer Review Data using Big Data: A Survey

Mugdha Jinturkar
Department of
Computer Engineering,
K.J. Somaiya College of
Engineering,Vidyvihar
Mumbai-77(India)

Pradnya Gotmare
Department of
Computer Engineering,
K.J. Somaiya College of
Engineering,Vidyvihar
Mumbai-77(India)

ABSTRACT
Rapid evolution in technology and the internet brought us to the era of online services. E-commerce is nothing but trading goods or services online. Many customers share their good or bad opinions about products or services online nowadays. These opinions become a part of the decision-making process of consumer and make an impact on the business model of the provider. Also, understanding and considering reviews will help to gain the trust of the customer which will help to expand the business.

Many users give reviews for the single product. Such thousands of review can be analyzed using big data effectively. The results can be presented in a convenient visual form for the non-technical user. Thus, the primary goal of research work is the classification of customer reviews given for the product in the map-reduce framework.

General Terms
Natural Language Toolkit (NLTK), Naïve Bayes Algorithm, SentiWordNet, Amazon

Keywords
Opinion Mining, Sentiment Analysis, Big Data, Data Visualization, Customer Reviews

1. INTRODUCTION
Almost 85% customers read online reviews before making a purchase. The providers would get feedback from the reviews which would help them for improvements in the upcoming products. Usually, reviews are given in text format. The single product has a number of reviews. It is hardly possible to read each review in detail. Many kinds of research showed pictorial representation is more effective and easy to understand rather than textual representations.

Existing work shows that various approaches are used for sentiment analysis founded on machine learning, a bag of words, natural language processing or even clustering. Though lots of work is done in the study domain, very few researchers are applied in big data framework to analyze mined opinions.

Traditionally, reviews are classified into two categories: Positive and Negative filtering out neutral reviews. But Neutral review plays the important role in decision making. If neutral reviews are taken into consideration, the effectiveness of results will improve. Moreover, Limitation of prior works includes complex data visualization like SentiCompass [1], OpinionSeer [13] and untrusted product review source like twitter data as twitter do not verify purchase done by the reviewer.

So, research work focuses on a novel approach to convert textual reviews into the visual representation, by adopting Hadoop environment for sentiment analysis to improve efficiency and to determine the attitude of mass towards the subject of interest.

The scope of research work does not include:

a. Consideration of emoticons

b. Work aims only to evaluate expressions about the product not the betterment of the product

The remainder of paper is organized as given below: Section 2 presents current knowledge including substantive findings, as well as theoretical and methodological contributions from existing work to a research work. Section 3 outlines the proposed system overview and process pipeline. Section 4 gives implementation requirement and details. While in Section 5, a summary of research work and future scope is mentioned.

2. RELATED WORK
Sentiment Analysis is a process applied on document level, paragraph level or sentence level to find out favored polarity that is positive, negative and neutral. Opinion mining is established in the domain of microblogging to RSS feeds, investor's choice to customer's choice, social issues to politics. Sentiment analysis aims to find out the opinion of mass towards the subject of interest.

There are various approaches for sentiment analysis:

1. Dictionary based approach:

The methods introduced in [17], [15] are built on opinion words that are commonly used in expressing positive or negative sentiment. This type of approaches uses SentiWordNet [23], tagged corpora with a positive score and a negative score along with part of speech tags. [16] proposed a novel lexicon based approach for opinion mining and also their own published lexicon dataset.

2. NLP-based approach:

In [21] Instead of classifying the whole document, Subject-oriented classification helps to improve precision. Thus, in this approach, using natural language processing they have identified the semantic relationship between subject terms and sentiment expression. Natural language processing assists in determining the structure of a sentence. [5] gives a brief idea about phases involved in such a kind of approach.

3.Machine Learning approach:

Pang & Lee, 2008 introduced machine learning approach which has two stages training stage and testing stage. In a training phase, sample dataset is classified using algorithms like Naïve-Bayes classifier, Support Vector Machine, Maximum Entropy model, etc. For domain-specific training data, efforts for manual tagging are necessary. Test dataset is then classified using trained model. 4.Ontology-based approach:

[10], Twitter which is micro-blogging website is used as a data source. Tweets related to smartphones are used as input data. Domain-specific ontology i.e. ontology for smartphones is built in the first phase. Each tweet is divided into a set of aspects relevant to the subject. In result, tweets are characterized by sentiment score as well as sentiment grade.

[9] gives a detailed survey of techniques which are used for sentiment analysis. Survey shows that sentiment analyzers are language reliant. It provides a comparative study of different techniques.

SentiCompass [1] is an interactive visualization of time-varying twitter data. It has obtained data using twitter API for two cases sports and political election. Work done on sentiment analysis and visualization of twitter data prior to SentiCompass has missing temporal data representation. SentiCompass Visualization includes TimeTunnel representation and Russell Circumplex‟s

Model [23]. Sentiment analysis of tweets is done by using ANEW (Affective Noun for English Words) [22] which have ratings for 1340 words. For deciding polarity of sentiment, Naïve Bayes Classifier [4] is used. Visualization is enriched with various interactive feature like hover, zooming, etc. Thus, SentiCompass overcome the problem of temporal data representation successfully.

Sentiment Analysis Using Big Data [3] proposed a framework which has two features:

a. Analyzing opinions in map-reduce environment

b. Lexicon based technique to extract neutral reviews and restricting reviews being classified into positive or negative.

The framework is applied to the customer reviews given on twitter. They considered reviews for “Google Glass” for demonstration. Usually, neutral reviews are ignored because the neutral text is not as important as clear positive or clear negative reviews are. But to achieve accuracy, Koppel, and Schler [18] shown that every polarity must be taken into consideration. Naive Bayes classifier is designed in the map-reduce environment. The framework identifies the polarity of sentiment, classifies into three categories (i.e. positive, negative and neutral) and then visualizes the result. Use of Hadoop provided the better capability of analysis.

In [11] they have done mining on book reviews to identify features of similar books and to compare user ratings for the same. They used Goodreads as a source of book reviews. On Goodreads, users can keep a log of books they have read, along with opinion about the book, expressed in terms of 5-star rating, and in the form of written reviews. The term frequency-inverse document frequency method (TF-IDF) is basic automated text mining technique. They marked frequently occurring words and produced a vector of frequency. Then calculated the weight of each term of each book. Then a global weight of each term is calculated. At last, Book similarity is calculated based on weights of tag word. Application of a hierarchical clustering technique is done in this case. Clusters were built up in consecutive rounds, by combining the two clusters with maximum similarity in each round. Then they have done cluster evaluation to compare ratings for similar kind of books. Thus, they proposed a method of attribute based mining from book reviews by identifying book features in the review text.

OpinionSeer [13] is an interactive visualization system which visually analyze a large collection of online hotel customer feedbacks. It uses www.tripadvisor.com as a source of customer reviews. OpinionSeer helps to assist travelers using various parameters for visualization. They use a new feature-based opinion mining technique to define the uncertainty in the review text.

The work is done in two domains:

a. Opinion Mining

b. Opinion Visualization

For opinion mining, they used a method proposed by Liu et. al [20] to extract feature level opinions from customer reviews. Using Subjective logic, they have done document level opinion mining. Major visual representation is OpinionWheel.Interactive features include brushing, linking, selection. Thus, In OpinionSeer, opinion extraction, analysis, and visualization are done. They achieved reliability in analysis and flexibility in visualization.

The represented approach in [2]is an analytical study on smartphones and they published dataset of thousands of remark on various smartphones. They used linguistic appraisal model. Also, they established an approach using natural language processing, opinion mining, and sentiment analysis. The focus is on aspect level opinion mining, building a platform which summarizes and qualifies experience feedback from reviews. To achieve maximum effectiveness, the use of a linguistic framework which focuses on appraisal in English [19] is required. The framework includes two main steps: the first one to create a knowledge base in order to extract relevant opinions and the second one to run syntactic analysis and search strategies.

Opinion mining of reviews involves:

a. Appreciation Extraction

An aspect-based opinion mining task is an appreciation extraction. They aim to extract three different expressions of appreciation [7]: explicit qualifier about an explicit aspect “The battery life is extensive”, “The problem is the sound quality”.

b. Affect Extraction

Affect is observation of positive and negative feelings. This method aims to extract two different expressions of affect [7]: affect as an “I am satisfied with this handset” which is positive, “I hate this handset” which is negative.

The knowledge base consists two resources: an aspect-lexicon which is a list of words and phrases defining the relevant aspects for an opinion mining task and a polarity lexicon [14] which gives us idea about the polarity of words. Lexical analysis, transformations of dataset and the syntactic analysis are done using OpenNLP [12].

Finally, custom knowledge base is referred in order to extract relevant speech elements related to appreciation or affect in the context of smartphone reviews. The approach is based on two key concepts: syntactic analysis and strategies to find appraisal patterns in parse trees [2]. As a result, platform that summarizes and qualifies user experience feedback is founded.

The study shows that every system has its own way for analysis like clustering. Also, they have different approaches for sentimentanalysis like lexicon based, NLP based. Visual representations also vary from search to search.

3. PROPOSED SYSTEM

In this section of paper, detailed introduction of the proposed system is given. Proposed system aims to analyze product reviews from an e-commerce website. The customer always prefers to read reviews before paying money to the service provider. But it is hardly possible to read all reviews in today's fast life. Also, every review may provide new information of product or feature of the product. So there is the probability of missing any important review given by consumer. So that, there is a need to identify the polarity of review i.e. whether it is positive, negative or neutral. Sentiment analysis is the best way to find out polarity. The consumer will be able to identify the polarity of review. once sentiment analysis results are received and he/she will take decision faster as efforts for reading reviews are reduced.

The Fig. 1 shows proposed system architecture. It gives an overview of the complete process pipeline. The system comprises of modules involved are:

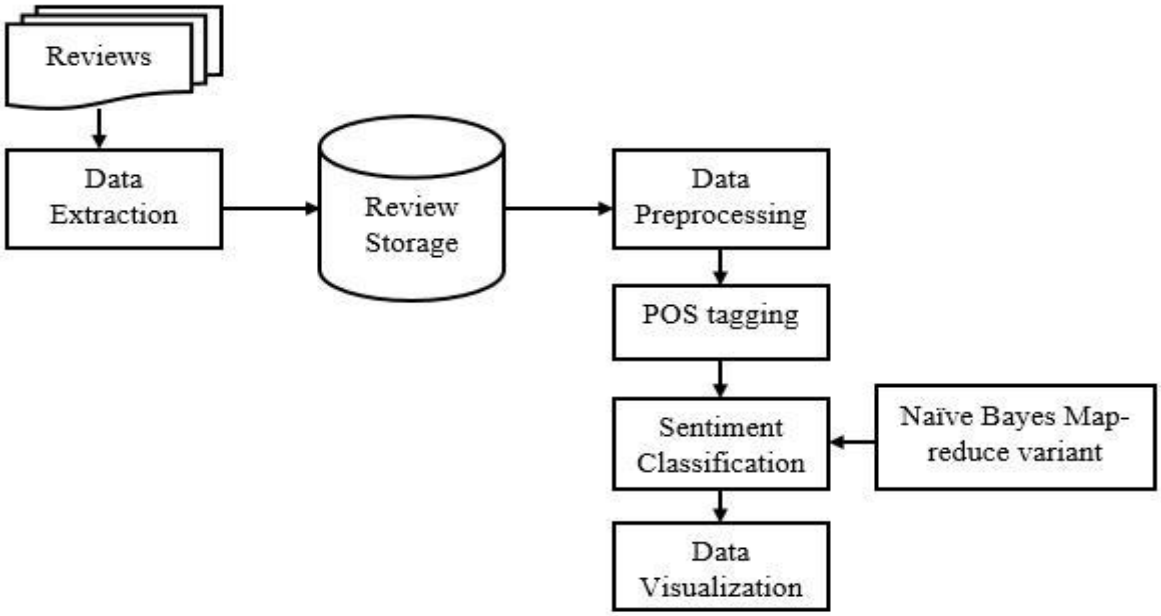

Fig.01 Proposed System Architecture

3.1 Data Extraction

In proposed system www.amazon.in is used as a source for data extraction. Reviews on Kindle are extracted using crawler implemented using Beautiful Soup library [25]. It automatically extracts all reviews on a single product using one seed URL. Then crawler will navigate through web pages to extract all the reviews. Extracted reviews are classified into two datasets which are used for training and testing. 70% of reviews are used for testing, rest of 30% reviews are used for testing.

3.2 Data Preprocessing

Data preprocessing includes proper fragmentation of data and cleaning of data. Here, in research work NLP preprocessing techniques [6] like the removal of stop words, chunking data, stemming etc. are used. Data preprocessing will lead us to robust data which has less noise. For data preprocessing, use of Natural Language Tool Kit (NLTK) library [26] implemented in python is considered. NLTK is a platform for natural language processing developed in python. Part of speech tagging [27] is also done using NLTK. It assists to profound sentence structure and interpret its meaning.

3.3 Sentiment Classification

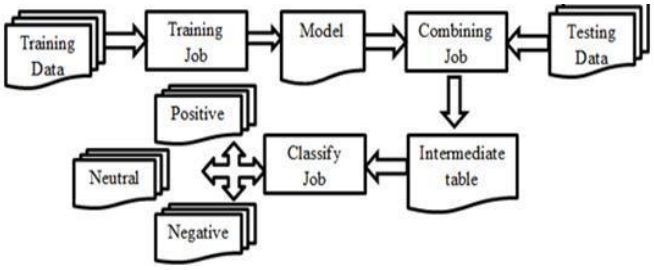

Fig. 2 Classification in Hadoop environment

The process of text classification is divided into two phases: Training phase and testing phase. In the training stage, the classification model is created using testing dataset. In the testing stage, the accuracy of classification is evaluated using classification module. The map-reduce environment is used to implement Naïve-Bayes classifier [3,8]. The processing is carried out as shown in Fig. 2.

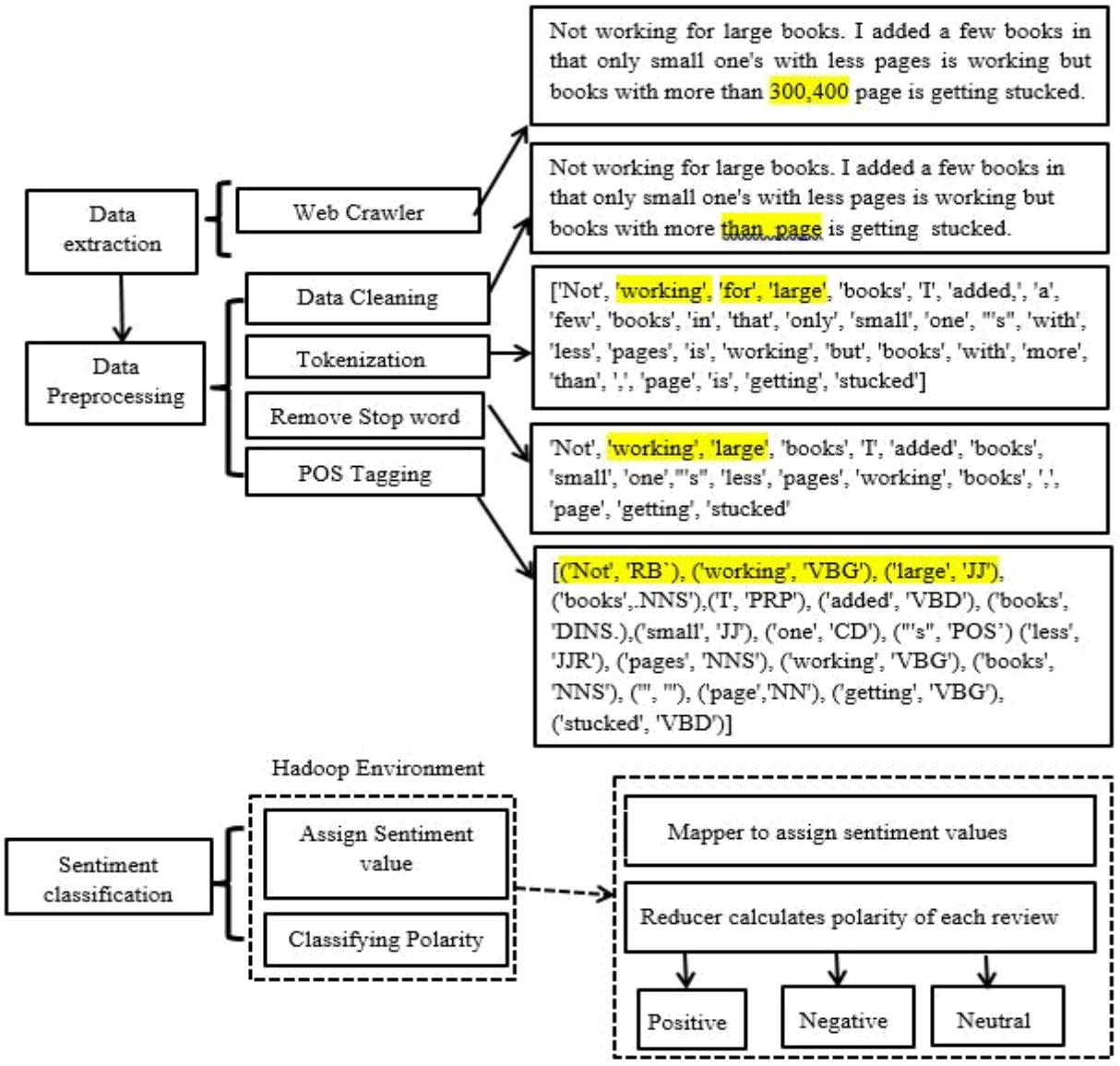

Fig. 3 Detailed processing overview

4. IMPLEMENTATION DETAILS

The data source to extract product reviews is www.amazon.in. All available reviews for a considered product are extracted and stored using web crawler. Data cleaning includes removal of digits, hyperlink, special characters from review text as they don't matter in sentiment evaluation. Contraction-expansion assists in the handling of negations. Using contraction expansion, terms like "don't" to "do not", "I'm" to "I am" are expanded.

Then tokenizer converts sentence into word tokens. Word tokens are then subject to stop word removal deals with common words like for, above, etc. Part of Speech(POS) tagging is applied to define grammatical tagging based on both its definition and its context. Data preprocessing is done using NLTK framework in python. Fig. 3 shows the illustrative example for the extraction and preprocessing module.

Sentiment classification is to be done in Hadoop environment. SentiWordNet is used to assign sentiment values using mapper job. SentiWordNet is a dictionary which has a positive, negative and objective score for each sentimental word. It has around 19000 adjectives and noun which expresses emotion.

Naïve-Bayes classifier is used to classify polarity using reducer job. Naïve Bayes is classical machine learning algorithm based on probability. This type of classifiers is highly scalable, and can handle a number of parameters effectively. Property of scalability of classifier leads to map reduce variant of the classifier. After training a dataset, a model is created by storing input and outputs of test dataset. Then combiner job will combine testing dataset with trained dataset and then results of final classification can be achieved.

5. CONCLUSION AND FUTURE WORK

The proposed method in this paper aims how to improve the quality of sentiment analysis on textual product reviews using Hadoop framework. Also, the methodology is based on training and testing will improve the accuracy of results of analysis. The focus is on the use of open source technologies mainly. However, proposed system has tremendous practical applications for both individual customer and service provider. The individual customer takes its benefit for decision making and service provider can take advantage to improve the quality of service as well as for new product design.

The partial results of proposed system are added into paper as the research work is in early stage of implementation. Thus, we conclude:

a. A proposed new approach using open source technologies to represent textual reviews in the form of visual representation.

b. NLP-based text classification is to be used to improve the effectiveness of analysis.

c. Application of map-reduce environment will help us to improve speed and reliability of analysis.

Future work includes:

a. Visualization of obtained results.

b. Aggregating reviews from two or more sites.

c. Feature Extraction from textual review data.

d. Consideration of Emoticons.

e. Application of data preprocessing in Hadoop environment.

6. ACKNOWLEDGEMENT

Every project is lead successfully because the efforts and the guidance of a number of people who have always given their helping hand to resolve problems. I sincerely appreciate the inspiration; support and instructions of all those people who have been taking efforts in making this project a success.

7. REFERENCES

[1] Wang, F.Y.; Sallaberry, A.; Klein, K.; Takatsuka, M.; Roche, M., "SentiCompass:Interactive visualization for exploring and comparing the sentiments of time-varying twitter data," Visualization Symposium (PacificVis), 2015 IEEE Pacific , vol., no., pp.129-133,14-17 April 2015

[2] Brisson, L.; Torrel, J.-C., "Opinion mining on experience feedback: A case study on smartphones reviews," in Research Challenges in Information Science (RCIS), 2015 IEEE 9th International Conference on, vol., no., pp.187-192, 13-15 May 2015

[3] Ramanujam, R.S.; Nancyamala, R.; Nivedha, J.; Kokila, J., "Sentiment analysis using big data," Computation of Power, Energy Information and Commuincation (ICCPEIC), 2015 International Conference on, vol., no., pp.0480-0484, 22-23 April 2015

[4] P. Gamallo and M. Garcia. Citius: A naive-bayes strategy for sentiment analysis on english tweets. In Proceedings of International Workshop on Semantic Evaluation 2014, pages 171–175, Aug 2014.

[5] R. S. Dudhabaware ; M. S. Madankar, "Review on natural language processing tasks for text documents," Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on, Coimbatore, pp. 1-5.doi: 10.1109/ICCIC.2014.7238427

[6] F. L. d. Santos and M. Ladeira, "The Role of Text Pre-processing in Opinion Mining on a Social Media Language Dataset," Intelligent Systems (BRACIS), 2014 Brazilian Conference on, Sao Paulo, 2014, pp. 50-54

[7] L. Zhang, W. Xu and S. Li, "Aspect identification and sentiment analysis based on NLP," Network Infrastructure and Digital Content (IC-NIDC), 2012 3rd IEEE International Conference on, Beijing, 2012, pp. 660-664.doi: 10.1109/ICNIDC.2012.6418838

[8] Bingwei Liu, E. Blasch, Yu Chen, Dan Shen and Genshe Chen, "Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier," Big Data, 2013 IEEE International Conference on, Silicon Valley, CA, 2013, pp. 99-104. doi: 10.1109/BigData.2013.6691740

[9] K. Ghag; K. Shah,"Comparative analysis of the techniques for Sentiment Analysis", Department of Information Technology, MET's Shah & Anchor Kutchhi Engineering College, Mumbai 400706, India","Advances in Technology and Engineering (ICATE), 2013 International Conference on","20130606","2013

[10] Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, Nick Bassiliades, Ontology-based sentiment analysis of twitter posts, Expert Systems with Applications, Volume 40, Issue10, August 2013, Pages 4065-4074, ISSN 0957-4174

[11] Lin E.,Shiaofen Fang; Jie Wang, "Mining Online Book Reviews for Sentimental Clustering," in Advanced Information Networking and Applications Workshops (WAINA),2013 27th International Conference, vol., no., pp.179-184, 25-28 March 2013

[12] J. Kottmann, B. Margulies, G. Ingersoll et al., "Apache opennlp. The apache software foundation," 2013.

[13] Yingcai Wu; Furu Wei; Shixia Liu; Au, N.; Weiwei Cui; Hong Zhou; Huamin Qu, "OpinionSeer: Interactive Visualization of Hotel Customer Feedback," in Visualization and Computer Graphics, IEEE Transactions on , vol.16, no.6, pp.1109-1118, Nov.-Dec. 2010

[14] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining." in AAAI fall symposium: common sense knowledge, vol. 10, 2010, p. 02.

[15] A. Neviarouskaya, H. Prendinger, M. Ishizuka, SentiFul: Generating a reliable lexicon for sentiment analysis, Proceedings of the affective computing and intelligent interaction and workshops (ACII 2009), 3rd international conference on affective computing and intelligent interaction and workshops, IEEE (2009), pp. 10–12 September 1–6arning (EMNLP-CoNLL) (pp. 1075–1083)

[16] Ding, X., Liu, B. and Yu, P. A Holistic Lexicon-Based Approach to Opinion Mining. Proceedings of the first ACM International Conference on Web search and Data Mining(WSDM'08), 2008.

[17] Kaji, N., & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In Proceedings of the joint conference on empirical methods in natural language processing and computational natural language le

[18] M. Koppel and 1. SchIer (2005) "The Importance of Neutral Examples for Learning Sentiment". In IJCAI

[19] P. R. R. White; J. R. Martin, The Language of Evaluation: Appraisal in English. London/New York: Palgrave Macmillan, 2005.

[20] Hu, M and Liu, B. "Mining and Summarizing Customer Reviews". Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 2004.

[21] Tetsuya Nasukawa, Jeonghee Yi, Sentiment analysis: capturing favorability using natural language processing, Proceedings of the 2nd international conference on Knowledge capture, October 23-25, 2003, Sanibel Island, FL, USA

[22] M. M. Bradley; P. J. Lang. "Affective norms for english words (ANEW): Instruction manual and affective ratings." Technical report, The Center for Study of Emotion and Attention, University of Florida, 1999.

[23] L. Feldman Barrett; J. A. Russell. "Independence and bipolarity in the structure of current affect." Journal of personality and social psychology, 74(4):967–984, 1998.

[24] SentiWordeNet , http://sentiwordnet.isti.cnr.it/

[25] Vineeth G. Nair. 2014. Getting Started with Beautiful Soup. Packt Publishing.

[26] Perkins, J. (2010) Python text processing with NLTK 2.0 Cookbook,Packt Publishing.

[27] Beatrice Santorin "Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)", University of Pennsylvania.