A Hybrid Approach for Aspect Based Sentiment Analysis on Big Data

A. Nandhini[1], G. Vaitheeswaran[2], Dr. L. Arockiam[3],

[1] M.Phil Scholar, Department of Computer Science, St. Joseph's College (Autonomous), Trichy-2, Tamilnadu, India
[2] Ph.D Scholar, Department of Computer Science, St. Joseph's College (Autonomous), Trichy-2, Tamilnadu, India
[3] Associate Professor, Department of Computer Science, St. Joseph's College (Autonomous), Trichy-2, Tamilnadu, India

---***---

Abstract - In the digital world, people post their reviews about the product, movies or anything else. When the data size increases in massive scale via posting reviews, it leads to the emergence of the big data. These reviews enable the people to extract the necessary information so that they can take decisions about the things. This process of analyzing the individual opinions conveyed in a piece of text is known as sentiment analysis also known as opinion mining. The aspect based sentiment analysis (ABSA) is a part of sentiment analysis, which extracts the feature of a product and gives the opinion. From the business standpoint, ABSA is crucial to cognize the needs of the marketplace. To perform effective Aspect Based Sentiment Analysis, advanced technologies and tools are required. This paper discusses the big data, sentiment analysis and Aspect Based Sentiment Analysis, and its issues such as aspect extraction and polarity measure. Aim of this study is to improve the accuracy of aspect and polarity extraction by using a hybrid approach.

Key Words: *Big Data, Sentiment analysis, aspect based sentiment analysis, aspect extraction.*

1. INTRODUCTION

The world population reached 7.28 billion in 2014. Also the data produced by the people too sharply increased [1]. From this statistics [2], there are 5 billion people having mobile phones, and 2 billion people are using internet. There is no standard definition for Big Data; simply we call it as large volume of data. The Big Data Commission at the TechAmerica Foundation offers the accompanying definition:

"Big Data is a term that describes large volumes of high-velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information" (TechAmerica Foundation, 2012) [3].

Characteristics of big data

A. Volume
B. Variety
C. Velocity

Sentiment analysis is a digital work to give any opinion, sentiments or subjectivity of the text or data. Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity [4].

There are 3 levels of sentiment analysis. They are

A. Document level
B. Sentence level
C. Aspect/ entity level

A. Document level

The document level aims to classify the whole opinion document as positive, negative or neutral.

B. Sentence level

In the sentence level, sentence by sentence evaluation is done to decide whether each sentence expressed a positive, negative, or neutral opinion.

C. Aspect/Entity level

Aspect refers to feature.

Aspect level was earlier called feature level (feature-based opinion mining and summarization). Objective of sentiment analysis: Given an opinion document d, discover all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) in d.

Sentiment analysis tasks

From the definition, we can specify the task in the sentiment analysis.

Task 1 (entity extraction and categorization):

Extract all entity expressions in d, and categorize or group synonymous entity expressions into entity clusters (or categories). Each entity expression cluster indicates a unique entity $e_{i.}$

Task 2 (aspect extraction and categorization):

Extract all aspect expressions of the entities, and categorize these aspect expressions into clusters. Each aspect expression cluster of entity e_i represents a unique aspect aij.

Task 3 (opinion holder extraction and categorization):

Extract opinion holders for opinions from text or structured data and categorize them. The task is analogous to the above two tasks.

Task 4 (time extraction and standardization):

Extract the time when opinions are given and standardize different time formats. The task is also analogous to the above tasks.

Task 5 (aspect sentiment classification):

Determine whether an opinion on an aspect a_{ij} is positive, negative or neutral, or assign a numeric sentiment rating to the aspect.

Task 6 (opinion quintuple generation):

Produce all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) expressed in document d based on the results of the above tasks [18].

2. RELATED WORKS

Avital et al in [7] described the characteristics of big data, and its issues and challenges including security privacy, trust and also discussed the technologies such as the Map Reduce, Hadoop to deal with the big data issues.

Liu et al in [8] described 3 levels of sentiment analysis and also discussed the problems of sentiment analysis and object identification, sentiment extraction, integration, feature extraction.

Long et al in [9] explored ABSA, a third level of sentiment analysis and discussed subtasks of ABSA including aspect term extraction, aspect term polarity, aspect category detection, aspect category polarity

Amani et al in [6] described 2 types of aspects namely implicit aspect and explicit aspect. Implicit aspect refers to the feature or Polarities indirectly.

E.g.1 this phone is very sleek and comfortable.

This example is indirectly refer to phone size. Therefore this type of aspect is called as implicit aspects.

Explicit aspect refers to the feature directly.

E.g.2 the picture quality of this phone is good

This example directly refers the picture quality of the phone. Here picture quality is the aspect.

Aspect grouping is basically grouped by the synonym of a word. Each category represents a unique aspect. Generally Wordnet or thesaurus dictionaries help to identify the group.

E.g.1 The battery life is good

E.g 2 this mobile battery is good.

The above 2 examples are grouped under same category 'battery'. And also the battery power comes under the 'battery' category.

Amani et al in [6] discussed 3 types or levels of polarity such as positive, negative and neutral

Amani et al in [6], discussed 3 types of polarity such as positive, negative and neutral

(i) Positive Polarity refers to positive comments about the aspect.

E.g.1 the voice quality of the phone is amazing

It gives the positive comment about the phone. Here the voice quality (audio) is the aspect.

(ii) ***Negative Polarity*** refers to negative comments of the aspect.

E.g.2 the zooming of the camera is bad

From the above example, the aspect zooming gives negative opinion of the camera.

(Iii) Neutral Polarity refers to both positive and negative polarity.

Cyril et al [5] explored 3 types of evaluation such as precision, recall, F-Measure.

Precision and recall are defined as follows:

$$Precision = \frac{|Extracted\ Aspects\ \cap\ True\ Aspects|}{|Extracted\ Aspects|}$$

$$Recall = \frac{|Extracted\ Aspects\ \cap\ True\ Aspects|}{|True\ Aspects|}$$

Sometimes F-measure is also used for evaluation of extracted aspects:

$$F - Measure = 2 * \frac{Recall * Precision}{Recall + Precision}$$

Aishwarya et al in [10] had taken an educational dataset. And proposed a new sentiment analysis algorithm called senti-score algorithm.

Deepak et al in [11] had taken the voting system datasets and evaluated using the supervised learning algorithm. Also for the sentiment analysis they used Random Forest classifier.

Soujanya et al in [12], proposed a novel rule based approach for the aspect and polarity measure using the pos tagging technique which identifies the Implicit Aspect Clue (IAC) and then maps the corresponding aspects.

koji et al in[13], had performed the word pair process using the two algorithms (TF-IDF) to extract the words related to the aspect.

Zhai et al in [14] proposed a constrained LDA algorithm to group the data into a similar part to improve the accuracy and also used probability aggregation and relaxation method for the calculation.

Lei et al in [15] used the double propagation algorithm and web page ranking algorithm HITS for aspect based sentiment analysis and also described the phrase pattern method to determine the accuracy in sentiment analysis.

Muhhamad et al [16], used the Feature Extraction and Linear Discriminant Analysis (LDA) algorithm for Preterm Birth Monitoring for the new born baby skin protection to find a linear transformation such that feature clusters are most separable after the transformation of the babies.

Qiu et al in [17] used the double propagation algorithm based on bootstrapping that needs an initial opinion lexicon to start the bootstrapping process.

3. PROPOSED WORK

The proposed work concentrates on the extraction accuracy of aspect and the polarity. In the proposed work, a new algorithm has been developed to improve the accuracy of the both extraction.

3.1 BASE PAPERS

3.1.1 Pos Tagging

In this paper the author used the pos tagging as one module, aspect extraction as another module, and then opinion words extraction as another module [6]. There was a feature dictionary to extract the features. And also they had the positive and negative seed list for opinion extraction. By using their proposed technique the aspect extraction were increased but the opinion extraction not increased. To resolve this problem, this study used another algorithm in [10].

3.1.2 Senti-Score

In this paper [10] the author discussed only about the sentiment words (opinion extraction). He proposed an algorithm called SentiScore algorithm.

The methodology involves the preprocessing, score board generation (SentiScore), classification process. By using the preprocessing can eliminate the un wanted text, which will help to improve the accuracy and the time consuming. He had taken the restaurant dataset to evolve. And then the score board generation as like +1,-1. 0 and so on, based on the sentiment words. Then the classification process is done by KNN algorithm, and also he explained about the KNN and the Naive bayes algorithms. The naïve bayes is perfect for the small amount of data to be processed but when the amount of data getting increased the accuracy getting low. So finally the KNN algorithm will suite for the large volume of data. By hybrid they both algorithms can get more accuracy.

3.1 PROPOSED FRAMEWORK

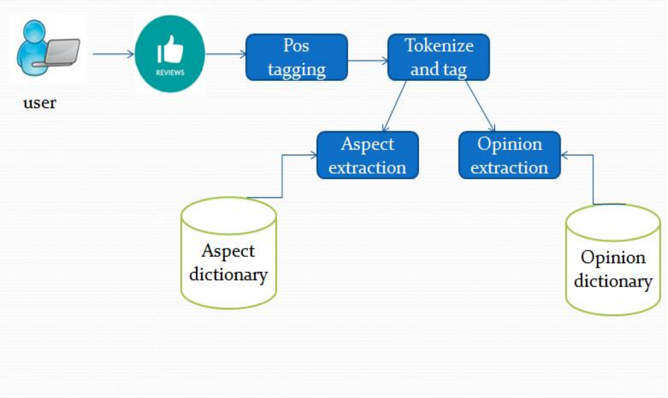

Diagram 1: Aspsen

All the users post their reviews in the social networks or in the Blogs. All the reviews are put into the POS tagging. The Pos tagging will split the sentence as tokenized and tag. The tokenized sentences will align the tag for the given tokenized words. The aspects will be extracted from the aspect dictionary, and the opinion words will be extracted from the opinion dictionary. The dictionaries are saved in .yml format in python; the programming codes will be written in python. The diagram 1 illustrates the proposed framework.

3.2 FLOW DIAGRAM:

The flow diagram describes the flow of algorithm. The reviews extracted from the web sites can be movies, product or anything else. The POS Tagger will separate the sentences i.e reviews for tokenizing and tagging. Diagram 2 describes the flow of the research work.

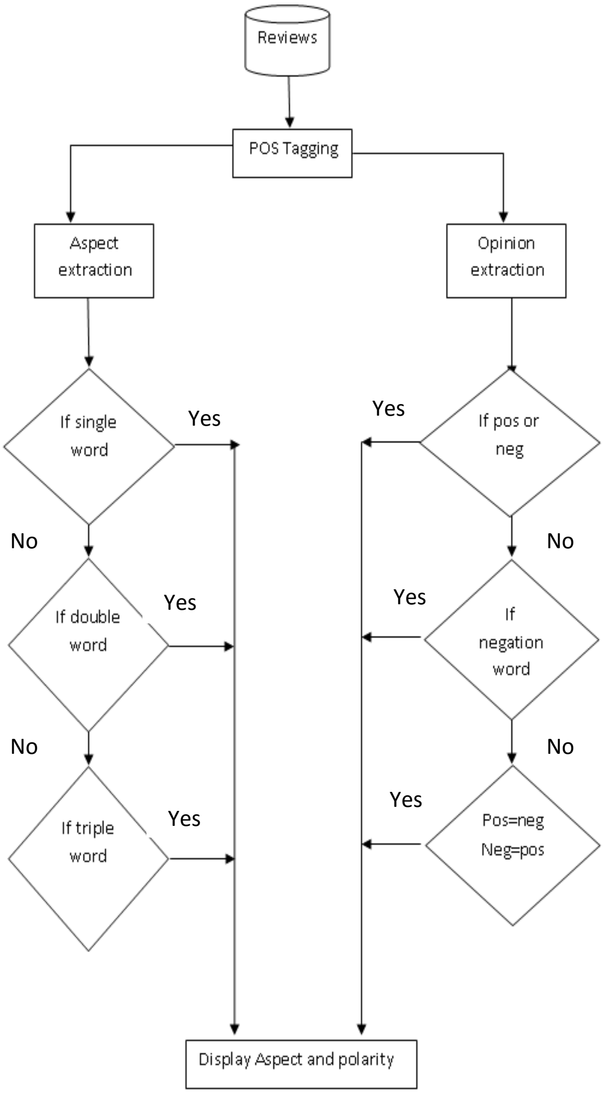

Diagram 2: Flow of Research

Table 1: Tokenizing and Tagging

Tag	Description
JJ	Adjective
JJR	Comparative adjective
JJS	Superlative adjective
LS	List item marker
NN	Noun, singular or mass
NNS/NNP	Noun, plural noun, singular
NNPS	Proper noun, plural
RB	Adverb
RBR	Comparative adverb
RBS	Superlative adverb
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund, or present participle
VBN	Verb, past participle
VBP	Verb, non-3^{rd}-person singular present
VBZ	Verb, 3^{rd}-person singular present

Table – 2: Example of POS Tagging

Tag	Sentences
[NN][VBZ][RB][JJ]	software is absolutely terrible
[NNS][VBP][JJ]	pictures are razor-sharp
[NN][VBZ][RB][JJ]	earpiece is very comfortable
[NN][VBZ][JJ]	sound is wonderful
[NNS][VBP][RB]	transfers are fast
[VBZ][JJ]	looks nice

3.2 PSEUDO CODE FOR PROPOSED WORK

```
Step 1: Extract every sentence S.
i.e for every sentence € $s_i$
Step 2: Remove the unwanted words by preprocessing.
Step 3: To tokenize and tag the words, want to put the sentence into the POS Tagger.
Step 3: The noun will be the aspect and the adjective will be the polarity of the sentence.
Step 4: The aspect can be a single word (unigram), double word (bigram), triple word (trigram).
            i.e $W_{i=}$ Unigram, Wi+1= Bigram, wi+2= Trigram
Step 5: Extract the aspect from the aspect dictionary.
Step 6: The opinion words will be extracted from the opinion lexicon.
Step 7: The opinion words can be positive, negative or neutral.
Step 8: The opinion words may have a negation word.
            i.e NOT
Step 9: The opinion word may have booster word.
            i.e Very
Step 10: Finally the aspect and the polarity values will be calculated.
```

Every sentence is represented as S, and the current sentence represented as $S_{i.}$ i may range from 1 to n (n number of sentences can be evaluated). To do the entire sentiment analysis task, first remove the unwanted words from the sentences using preprocessing. The preprocessing can be done by any tool like Stanford or anything else. Then the sentences will be tokenized and tagged by the POS Tagger. The POS Tagger will take the frequent noun alone. Typically the frequent tag will be the noun which represents the aspect and the adjective which represents the polarity i.e opinion word. The aspect cannot be a single word (unigram); it can be double (bigram) or triple words (trigram). The entire features will be extracted from the feature dictionary that may be domain based..

The opinion words extracted from the opinion lexicon can be positive, negative or neutral. And also the opinion can contain a negation word (NOT), which will give opposite meaning to the polarity. For example the mobile is not good. It has the word good, which will give the positive polarity, but before that it has the word net. So the meaning will change as bad, which represent the negative polarity. And also the opinion word can contain the booster word very, which will give more value to the feature. Finally all the feature and polarity are collected and based on that decision will be taken as positive, negative or neutral.

4. CONCLUSIONS

Nowadays handling of Big Data is a big challenge due to huge volume of the data. From the business viewpoint, ABSA is crucial to cognize the needs of the marketplace. Customers play a vital role in facilitating business intelligence through their reviews. Based on such reviews companies can improve the performance of their products. But extracting only the necessary information from the reviews is a big challenge, because of huge volume of data. This process of analyzing the data comes under the sentiment analysis and extracting the aspects of the product comes under aspect based sentiment analysis. That is extracting the feature and polarity measure about the product. By implementing the proposed tools and technologies, the Big Data handling becomes easier.

REFERENCES

[1] Worldometers, "Real time world statistics,"2014, http://www.worldometers. Info/world-population.

[2] D.Che, M.Saffron, and Z.Peng, "From Big Data to Big Data Mining: challenges,issues,and opportunities," in Database Systems for Advanced Applications,pp. 1–15,Springer,Berlin,Germany, 2013.

[3] Jean Yan, U.S. General Services Administration "Big Data, Bigger Opportunities", April 9, 2013.

[4] Walaa Medhat- School of Electronic Engineering, Canadian International College, Cairo, Campus of CBU, Egypt, Ahmed Hassan, Hoda Karachi-Ain Shams University, Faculty of Engineering, Computers & SystemsDepartment, Egypt, "Sentiment analysis algorithms and applications: A survey", 27 May 2014.

[5] Cyril Goutte and Eric Gaussie, "A Probabilistic Interpretation of Recall and F-score, with Implication for Evaluation", Proceedings of the European Colloquium on IRResarch (ECIR'05), LLNCS3408 (Springer), pp.345-359.

[6] Amani K Samha, Yuefeng Li and Jinglan Zhang," Aspect-based opinion extraction from customer reviews", international journal of Computer Science & Information Technology (CS & IT), 2014.

[7] Avita Katal, Mohammad Wazid, R H Goudar," Big Data: Issues, Challenges, Tools and Good Practices", IEEE, 2013.

[8] Bing Liu, "sentiment analysis and opinion mining", Mor gan Claypool Publishers, 2013.

[9] Long, Chong, Jie Zhang, and Xiaoyan Zhu. A review selection approach for accurate feature rating estimation . In Proceedings of coling 2010: Poster Volume, 2010.

[10] Aishwarya Mohan, Manisha. R, Vijayaa. B, Naren. J, "An Approach to Perform Aspect level Sentiment Analysis on Customer Reviewsusing Sentiscore Algorithm and Priority Based Classification", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4145-4148.

[11] Deepak Kumar Gupta, Asif Ekbal,"IITP:Supervised Machine Learning for Aspect based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 319–323.

[12] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, Alexander Gelbukh, "A Rule-Based Approach to Aspect Extraction from Product Reviews", http:// creativecommons. org/ licenses/by/4.0

[13] Koji Yatani, Michael Novati, Andrew Trusty, and Khai N. Truong, "Analysis of Adjective-Noun Word Pair Extraction Methods for Online Review Summarization", Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 2012.

[14] Zhongwu Zhai , Bing Liu , Hua Xu , Peifa Jia, "Constrained LDA for Grouping Product Features in Opinion Mining".

[15] Lei Zhang, Bing Liu, Suk Hwan Lim, Eamonn O'Brien-Strain, "Extracting and Ranking Product Features in Opinion Documents" 2009.

[16] Muhammad Naufal Mansor, Sazali Yaacob, Hariharan Muthusamy, "PCA- Based Feature Extraction and LDA algorithm for Preterm Birth Monitoring", International Journal of Soft C omputing And S oftware Engineering (JSCSE), 2011.

[17] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen, "Opinion Word Expansion and Target Extraction through Double Propagation",Association for Computational Linguistics,2011.

[18] Ku, Lun-Wei, Yu-Ting Liang, and Hsin-Hsi Chen. opinion extraction, summarization and tracking in news and blog corpora . In Proceedings of aaai-caa W '06. 2006.