Jenifer Jothi Mary A *et al*, International Journal of Computer Science and Mobile Applications,
Vol.5 Issue. 12, December- 2017, pg. 1-7 **ISSN: 2321-8363**
Impact Factor: 5.515

Aspect based Parsing for Sentiment Analysis in Big Data

Jenifer Jothi Mary A[1], Arockiam L[2]
[1]Doctoral Research Scholar, Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamilnadu, India, ajenifer.jothi@gmail.com

[2]Associate Professor, Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamilnadu, India, larockiam@yahoo.co.in

Abstract

Today online reviews become enormously valuable sources for mining customers' opinions on services and products. Extracting these opinions from these reviews and hulling the gold bar out of them are hard-won task. Though it is a herculean, it has great crunch on the decision making process of the companies and consumers. This is the reason for sentiment analysis to be a crowd-pleasing topic of research. There are many techniques proposed for improving the accuracy of the sentiment analysis using parts-of-speech (POS) approach. But the authors present an argument that POS tagging has time and space complexity in sentiment analysis and proposed a novel algorithm, Aro_Jen to enhance the performance of sentiment analysis without POS process. Aspect and sentiment based lexicons are used for experimental analysis to prove the claim of the time and space complexity of POS tagging process in sentiment analysis with twitter dataset. Result of this research has the potential of being successfully applied to eliminate POS-tagging process in other text classification problems.

Keywords: Aspect based parsing, Sentiment Analysis, Lexicon checking, Mob_Lex, Afinn, Big data

1. Introduction

Social networking sites have become more popular now-a-days as they provide platforms to its users to express their opinions on various topics such as products, brands and companies without any constraints and biases. These sites are the ideal sources for collecting customer opinions and developing the market intelligence. Moreover analyzing and summarizing these unstructured, large-scale opinions will offer useful knowledge to both companies and customers for making better decisions which affect the business outcome.

Sentiment analysis is the process of identifying and classifying opinions uttered in a piece of text, in order to determine whether the customer's attitude is positive, negative or neutral towards a particular topic, product, etc. In order to facilitate accurate sentiment analysis in user-generated contents, this paper introduces a domain specific aspect based parsing for sentiment analysis. It attempts to explain the un-necessary works done in POS Tagging in aspect based sentiment analysis and proposes a novel architecture for performing aspect based sentiment analysis without POS Tagging.

2. Background

Lexical parsing is a language processing technique, used to convert the series of characters into a sequence of tokens. It is very much useful in calculating sentiment score from online reviews and tweets. The process of aspect based sentiment analysis is depicted in Figure 1.

Lexical parsing breaks the given sentence S into a series of tokens, by removing all whitespace or comments in the source. Next, syntactic parsing is performed by applying the Part-of-Speech (POS) tagging or

©2017, IJCSMA All Rights Reserved, www.ijcsma.com

grammatical rules by which sentences are constructed. POS Tags such as adjective, adverb, noun and verb are applied on the words for making word extraction easier. The next task is checking for aspect words. If any aspect word is present, the lexical parser extracts adjective, adverb, noun and verb using 5-gram forwards and backwards searching from the aspect position [1].

Two important sub-tasks, namely, subjectivity and sentiment classifications are performed on each sentence *S*. If *S* is subjective then the parser will determine whether it expresses a positive, negative or neutral opinion, otherwise, it will be an objective sentence and it would be removed. These sub-tasks are useful to categorize sentences that have no opinions (objective sentences).

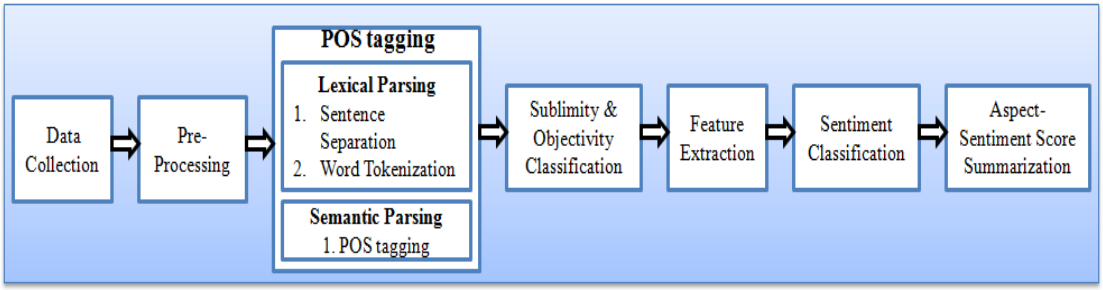

Figure 1 Processes of Aspect based Sentiment Analysis

3. Related Research

Richa Sharma et al. proposed an aspect based opinion mining system for classifying the reviews [2]. It extracted the feature and opinions from the sentences and classified them as positive, negative and neutral. Moreover negation was perfectly handled by the proposed system. A dictionary based technique of the unsupervised approach was adopted to determine the semantic orientation of the sentences. Experimental results expressed the effectiveness of the system.

Edison Marrese-Taylor et al. extended the aspect-based opinion mining approach proposed by Bing Liu [3]. In the proposed system, NLP-based rules were applied for subjective and sentiment classification at the aspect-level. Opinion visualization and summarization were entailed and made available to the online users. Generic architecture for an aspect-based opinion mining was developed, created a prototype and analyzed the opinions from TripAdvisor. The explicit aspect expressions used non-extended approach to extract reviews. It improved the accuracy and recall for subjective and sentiment classification.

Alexander Hogenboom et al. assessed various methods for supporting lexicon-based sentiment analysis [4]. The proposed system translated text into a reference language for sentiment lexicon and subsequently analyzed the translated text. The researchers considered mapping sentiment scores from a semantically enabled sentiment lexicon to a new target sentiment lexicon and traversed relations between language-specific semantic lexicons. Moreover the authors created a target sentiment lexicon by propagating sentiment of seed words in a semantic lexicon for the target language. It yielded a significant performance improvement in terms of accuracy and F-score. Depending on the seed set of sentiment-carrying words, the proposed work had language-specific dimension and outperformed the semantic baseline.

Petra Kralj Novak et al. presented a sentiment lexicon for emoji called as Emoji Sentiment Ranking and drew the sentiment map of the 751 most frequently used emojis [5]. The authors used 83 human annotators to label over 1.6 million tweets in 13 European languages. The significant differences of the sentiment distribution of the tweets with and without emojis were analyzed. The researchers observed that there were no significant differences in the emoji rankings between the 13 languages and also proposed Emoji Sentiment Ranking independent of European languages for automated sentiment analysis.

G. Wang et al. proposed an enhanced Random Subspace method, POS-RS, for sentiment classification based on part of speech analysis [6]. The proposed POS-RS employed content lexicon subspace rate and function lexicon subspace rate to control the balance between the accuracy and diversity of base learners. A detailed comparison with the performance of ensemble methods was provided by the researchers. Ten sentiment datasets were investigated to verify the effectiveness of the proposed method. The performance of ensemble learning methods was influenced not only by the diversity but also by the accuracy of base learners.

©2017, IJCSMA All Rights Reserved, www.ijcsma.com

Empirical results revealed that POS-RS achieved the best performance through reducing bias and variance simultaneously compared to the Support Vector Machine and yielded the lowest bias among all the ten datasets. Results illustrated that POS-RS could be used for sentiment classification and had the potential of being successfully applied to other text classification problems.

4. Motivation

Today analyzing the social networking sites reviews in a large-scale provide useful knowledge to both companies and customers for making better decisions. Business companies can improve and adjust their market strategies by analyzing their customers' opinions towards their brands and products. But aggregating, organizing, analyzing and summarizing these large amount of opinions in real time is much costly. Moreover, it is a hard-won task because the reviews are in informal languages with many short cuts, symbols and emojis. Though they convey rich information but detecting and annotating these informal reviews are laborious and troublesome. Thus, many proposed sentiment analysis systems use POS tagging to detect and tag the reviews as noun, adjective, verb, adverbs with the help of English language lexicons. Second, the online reviews are very casual and noisy. So, obtaining results from high-quality POS-tagging is a herculean task. Finally, many traditional words are used differently with different aspects in the micro-blogging scenario. So, it is an unyielding task to use POS tagging in sentiment analysis which has less impact on the aspect based sentiment analysis. To overcome these limitations, there is a need to develop a new technique which eliminates the POS tagging in sentiment analysis process without compromising the accuracy level of the sentiment analysis. Thus, this paper proposes a Aro_Jen tagging technique for the aspect based sentiment analysis.

5. Objective

The main objective of this paper is to propose a novel technique for aspect based sentiment analysis without POS tagging. The sub-objectives are defined as follows:

- To propose an effective new technique for aspect based sentiment analysis to eliminate the parsing of sentences and words in a document.
- To design a unified methodology to reduce the time complexity of the sentiment analysis system by phase out the POS tagging.
- To build an aspect based lexicon for improving the accuracy of sentiment analysis.

6. Proposed Work

The architecture of the proposed Aspect based sentiment analysis is given in Figure 2. It is proposed to eliminate the POS tagging and tokenization steps of the normal sentiment analysis. The forthcoming sub-sections describe the proposed work in detail.

- *Data collection:* This module collects the tweets of Redmi phone from the Twitter. A python crawler is used to extract the tweets which parse XML code for the tweet collection. To crawl twitter data, the user has to create a twitter API account. The extracted tweets are stored in JSON file and converted into csv format, which is used as the input to the aspect based sentiment analysis system.

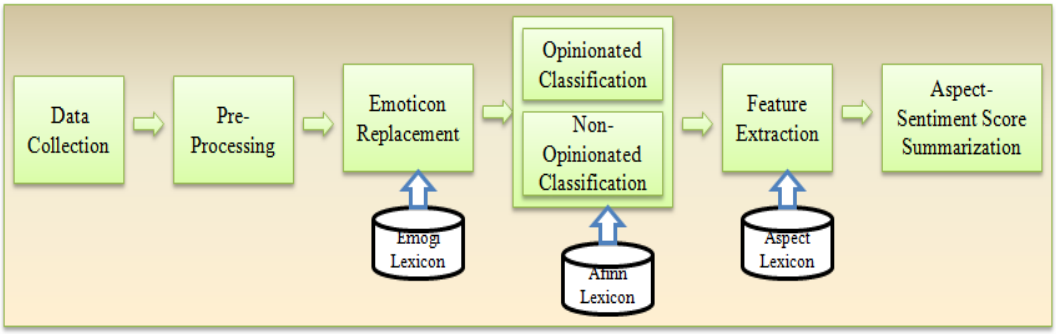

Figure 2. Proposed Architecture for Aspect based sentiment analysis

©2017, IJCSMA All Rights Reserved, www.ijcsma.com

- *Data Pre-processing:* When the user posts tweet, total of 3092 different information are generated. From these many fields, only the review field is selected for the pre-processing. This step removes unnecessary characters, hash tags and symbols to avoid the processing overhead.
- *Emoticon replacement:* The emoticons are icon / symbol and it is represented as ASCII in tweets. They are also called as *smileys.* It is formed by user by creative use of letters, numbers and punctuation symbols. Mostly they represent facial features but not at all the time. Emoticons are designed to add flavor to plain text. There are 2242 emoticons identified in the Twitter dataset [7]. For example a simple punctuation convey surprise ! , :-) - happiness and joy, :-(- sadness , :-D - laughter , ;-) – cheekiness.
- *Review Classification:* All the reviews in a document may or may not poses sentiment words. In such case, the proposed Aro_Jen procedure given in Figure 3 is used to classify the reviews as opinionated and non-opinionated without performing POS tagging. AFINN sentiment lexicon is used to perform this classification. Lexicon is a collection of words with their respective scores. Normally, the sentiment analysis is performed by considering only the sentiments.

Procedure: Aro_Jen()

Step 1: Read the Tweets
Step 2: Read the sentences using (.) separator
Step 3: Feed the Aspect and Sentiment Lexicons (Mob_Lex, Afinn)
Step 4: Select the aspect word and check for the corresponding sentiment
Step 5: If sentiment present, calculate the aspect based sentiment score
Step 6: Print the Aspect – Sentiment Report

Figure 3. Procedure Aro_Jen()

- *Feature – Sentiment extraction:* Feature extraction is nothing but aspect extraction. Aspects are extracted from the collected dataset using Aspect Lexicons that are specially developed for the domain indented. In this research, a Mobile Lexicon (Mob_Lex) is created with set of words that are frequently used to search for mobile phones. This lexicon is used for further analysis. If any word of the selected tweet have a match with Mob_Lex then, it immediately checks the sentiment lexicon Afinn [8], for related sentiment word for finding the polarity of an aspect. If sentiment word also is present in the same sentence of the tweet, then the polarity score of the word is collected from the Afinn and aspect based polarity is calculated for that particular aspect. This process is carried out for all the sentences of the dataset and polarity score for each aspect of that particular product (Mobile Phone) is calculated.
- *Aspect-Sentiment Score summarization:* Calculated aspect –sentiment scores are summarized for each aspect and the report is generated for that product.

7. Real Time Illustration

The real time illustration of the proposed algorithm is explained in this section.

Example Review 1: *Phone color is white. Its color is good.*

Let the given review be R1. The experimental analysis is presented in Table 1.

Table 1. Experimental analysis Aro_Jen()

Steps	Process	Action carried out
Step 1	Read the Tweets	Review R1 is read **R1:** *Phone color is white. Its color is good*
Step 2	Read the sentences using (.) separator	R1 is separated into two sentences S1 and S2 **S1:** *Phone color is white.* **S2:** *Its color is good.*

©2017, IJCSMA All Rights Reserved, www.ijcsma.com

Step 3	Feed the Aspect and Sentiment Lexicons (Mob_Lex, Afinn)	Mob_Lex and Afinn is feed in
Step 4	Select the aspect word and check for the corresponding sentiment	**S1:** Aspect word *color* is found but no sentiment word. **S2:** Aspect word *color* with sentiment word *good* is found.
Step 5	If sentiment present, calculate the aspect based sentiment score	Aspect based sentiment score for the word good (+3) is calculated
Step 6	Print the Aspect – Sentiment Report	Steps 4 and 5 are repeated for the whole dataset and finally all the score of a particular aspect is added up and report is generated using the formula $ASR(A_i) = \sum s_i$ Where, ASR is Aspect based Sentiment Report A_i – No. of aspects present in the dataset s_i – No. of sentiment words present for a particular aspect A_i in the dataset

8. Result and Discussion

In this section, the researcher tries to prove that the POS tagging is time consuming and having space complexity in aspect based sentiment analysis.

For the experimental analysis, tweets are crawled using an application designed in Python and NLP tool is used to pre-process the collected tweets for further processing. After pre-processing, POS tagging is applied on the cleaned tweets to prove the POS tagging process is not necessary for aspect based sentiment analysis.

Claim 1: *The POS tagging have space complexity.*
Proof:

In POS tagging, the review document is separated into sentences and tokenized as words. After the word tokenization, the NLP comparison is carried out [9]. So, it is necessary to store all the sentences and words of the review for further processing which requires more memory space. This huge amount of memory usage is overcome by the proposed architecture by using aspect based sentiment analysis without POS tagging. A comparative analysis is presented in Figure 4 for the memory requirement of POS tagging.

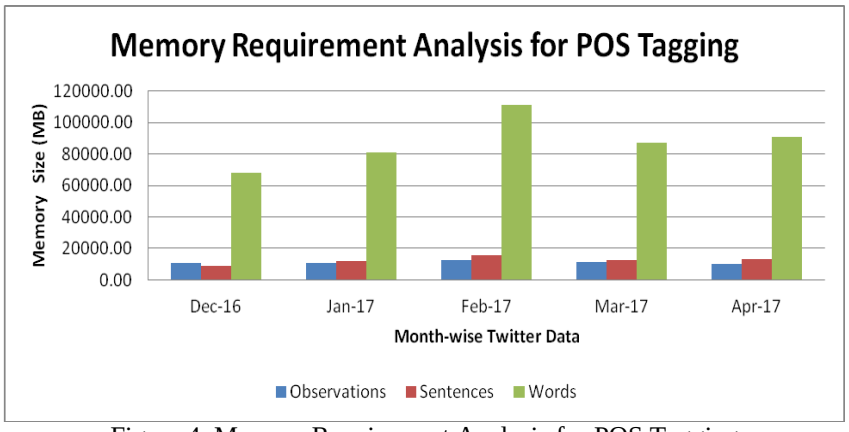

Figure 4. Memory Requirement Analysis for POS Tagging

Claim 2: *The POS tagging have time complexity*
Proof:

Normally, POS tagging is applied on every words in the review. If the reviews have good number of sentences, it takes more time to complete the process [10]. And POS tagging takes more execution time, almost triple the time of the other while comparing with the Aro_Jen tagging sentiment analysis. This can be avoided by

©2017, IJCSMA All Rights Reserved, www.ijcsma.com

bypassing the POS tagging procedure. The time taken for aspect based sentiment analysis with and without POS tagging is depicted in Figure 5.

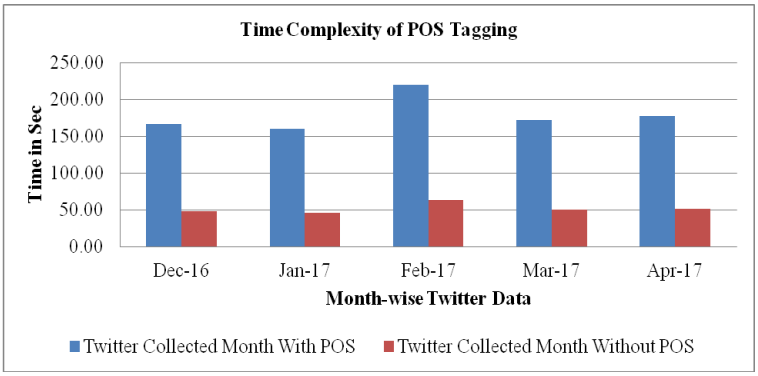

Figure 5. Time Complexity of POS Tagging

Thus it is proved that the task of POS tagging has time, space complexity in aspect based sentiment analysis. To overcome such issues, a new architecture and algorithm is proposed in this research work and it is proved.

9. Conclusion

In recent times, the sentiment analysis is becoming relevant to the current issues and aspect based sentiment analysis has become a research issue. In this paper, a novel architecture is proposed for aspect based sentiment analysis without parts-of-speach (POS) processes. The reason for eliminating POS tagging is presented and the proposed algorithm is experimentally proved. A proposed algorithm Aro_Jen is implemented and analyzed using the R tool. The derived results show that the proposed Aro_Jen algorithm not only requires less memory for storing intermediate results of the process but also takes less execution time. By using a aspect based lexicon (Mob_Lex) the number of comparison is considerably reduced in the process of aspect based lexicon checking. Thus, the proposed architecture reduces time and memory complexities and improves the lexicon checking.

References

[1] Singh, V. K., et al., 2013, Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. 2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed sensing (iMac4s) , IEEE.

[2] Richa Sharma,Shweta Nigam and Rekha Jain, 2014, "Mining of Product Reviews at Aspect Level", International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.3, DOI:10.5121/ijfcst.2014.4308 87

[3] Edison Marrese-Taylor, Juan D. Velásque, Felipe Bravo-Marquez, 2014, "A novel deterministic approach for aspect-based opinion mining in tourism products reviews", Expert Systems with Applications (Elsevier), Volume 41, No. 17, pp. 7764-7775, http://dx.doi.org/10.1016/j.eswa.2014.05.045

[4] Alexander Hogenboom, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, 2014, "Multi-lingual support for lexicon-based sentiment analysis guided by semantics", Decision Support Systems (Elsevier), Volume 62, pp. 43–53. http://dx.doi.org/10.1016/j.dss.2014.03.004.

[5] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, Igor Mozetič, "Sentiment of Emojis", PLoS ONE, Volume 10, No. 12, pp. 1-22. doi:10.1371/journal.pone.0144296 .

©2017, IJCSMA All Rights Reserved, www.ijcsma.com

[6] Gang Wang, Zhu Zhang, Jianshan Sun, Shanlin Yang, Catherine A. Larson, 2015, "POS-RS: A Random Subspace method for sentiment classification based on part-of-speech analysis", Information Processing and Management (Elsevier), Volume 51, No. 4, pp. 458-479, http://dx.doi.org/10.1016/j.ipm.2014.09.004

[7] http://datagenetics.com/blog/october52012/index.html

[8] http://www2.imm.dtu.dk/pubdb/views/publicationdetails.php?id=59819

[9] G. Vaitheeswaran and L. Arockiam, 2016, "Lexicon Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets", IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555, Vol.6, No3, pp. 33-38

[10] Muhammad Abdul-Mageed, Mona Diab, Sandra Kubler, 2014, "SAMAR: Subjectivity and sentiment analysis for Arabic social media" Computer Speech & Language, Volume 28, No. 1, pp. 20–37.

A Brief Author Biography

Ms. A. Jenifer Jothi Mary - She is a Part time research scholar and currently working as an Assistant Professor in the department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli. She has 5 years of Teaching experience and 3 years of Research experience. She has received her UG, PG and M.Phil degree from Bharadhidasan University, Tiruchirapalli, and B.Ed from Tamilnadu Teacher's Education Univeristy. Now she is pursuing her doctoral degree in the same university. Her area of research is Big Data Sentiment Analysis. She has also acted as a Resource person for the Short courses and State Level Workshop. She has presented papers many in National and International Conferences and also attended many workshops, Seminars and Training programs. She has published 4 papers in reputed journals with good impact factor.

Dr. L. AROCKIAM is working as Associate Professor in the Department of Computer Science, St.Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu. He has 29 years of experience in teaching and 20 years of experience in research. He has published 284 research articles in the International / National journals and conferences. He has also presented research articles in Italy, Indonesia and Malaysia. He has chaired 88 technical sessions and delivered invited talks in National and International Conferences. He has co-authored books on "Success through Soft Skills", "Research in a Nutshell", "Object Oriented Programming with C#.NET" and "WEKA: A Practical Guide to Beginners". His research interests are: Internet of Things, Cloud Computing, Big Data, Data Mining, Software Measurement, Cognitive Aspects in Programming, Web Services and Mobile Networks. Contact: 9443905333 Website: www.arockiam.in

©2017, IJCSMA All Rights Reserved, www.ijcsma.com