Real-Time Aspect-Based Sentiment Analysis on Consumer Reviews

Jitendra Kalyan Prathi, Pranith Kumar Raparthi and M. Venu Gopalachari

Abstract The rise of e-commerce websites, as new shopping channels, led to an upsurge of review sites for a wide range of services and products. This provides an opportunity to use aspect-based sentiment analysis and mine opinions expressed from text which can help consumers decide what to purchase and businesses to better monitor their reputation and understand the needs of the market. Aspect-based sentiment analysis (ABSA) is a technique aimed to foster research beyond sentence or text-level sentiment classification. The goal is to identify opinions expressed about specific entities (e.g., laptops) and their aspects (e.g., price, performance, build quality, etc.). There exist very few techniques which can generate such results based on customer ratings, however usually for a limited set of pre-defined aspects and not from free-text reviews. The other challenge in this process is cold start problem because of the lack of enough review data for a product. In this paper, a methodology is proposed to automatically compute sentiments of dynamic aspects from user-generated reviews from the web scraping from multiple sources to overcome the cold start problem. Therefore, this methodology is devising a better solution for understanding sentiments in e-commerce than existing methods.

Keywords Aspect-based sentiment analysis · Data integration · Cold start · Web scraping

J. K. Prathi (✉) · P. K. Raparthi · M. V. Gopalachari
Department of CSE, Chaitanya Bharathi Institute of Technology, Hyderabad, India
e-mail: kalyanjithendra27@gmail.com

P. K. Raparthi
e-mail: praneeth970@gmail.com

M. V. Gopalachari
e-mail: venugopal.m@cbit.ac.in

K. S. Raju et al. (eds.), *Data Engineering and Communication Technology*,
Advances in Intelligent Systems and Computing 1079,
https://doi.org/10.1007/978-981-15-1097-7_67

1 Introduction

In recent times, online reviews are very common to be found in lots of websites across the Internet, one particular kind is e-commerce reviews. People who shop online use reviews to get a brief explanation of the products or any information they need; however, preferences of the readers differ from one another. Some would like to find a product of specific brand, while others would be more interested in aspects such as quality, price, service, etc. On the other hand, the reviews written by different customers contain preferences of the respective reviewer. Therefore, readers will have to spend more time to go through the content and understand the opinions expressed on products. It will be of great use to customers if the websites provide the product with rating based on aspect categories rather than an overall rating. However, only few websites such as Amazon and Flipkart provide such rating system at present. To overcome this challenge, an automatic rating generator based on sentiment for each aspect is needed, where the technique to achieve this is aspect-based sentiment analysis [1]. Retrieving information from a user-generated content, particularly retrieving the sentiment in it requires the use of specialized NLP techniques [2]. Added the task of categorizing into aspects and retrieving sentiment for each aspect has proved to be quiet difficult. The user-generated reviews are of high value to the organizations as well as customers. Cold start is another challenge that degrades the quality of the information retrieval. Cold start refers to the new or less data in terms of size or users though there are several alternatives to remedy cold start problem such as using demographic information [9].

In this paper, an automated information retrieval system has been proposed that generates the sentiment analysis report on real-time reviews. The proposed system uses natural language processing techniques to mine the aspects and supervised learning techniques for sentiment classification. This system collects the name of a product as input and then uses spiders to crawl popular e-commerce websites and scrape the reviews from them. Then, the proposed methodology automatically extracts the potential aspects and corresponding entities and finally determines the sentiments expressed. These entities are matched to its aspect, and the system will measure the value of positive and negative sentiments for each aspect. The experiments conducted on the proposed system shown significant improvement in retrieving the information from multiple sources.

Rest of the paper is organized as follows: Sect. 2 contains the related work of proposed system, Sect. 3 contains the architecture and essential steps of automated ASBA, Sect. 4 describes results and discussions of proposed system, and Sect. 5 concludes research work with future directions.

2 Related Work

ABSA is a complex technique which is usually split into aspect extraction and sentiment analysis sub-tasks. Previous approaches to aspect extraction framed the task as a multiclass classification problem, and it relied on techniques with large corpora, e.g., named entity recognizer, parsing, semantic analysis, bag-of-words, as well as domain-dependent ones, such as word clusters [11]. The previous sentiment analysis approaches have used different classifiers with a wide range of features based on n-grams, POS, negation words, and a large array of sentiment lexica. We observed there is no technique which can automatically extract aspects from raw text without any previous information about the aspects. Sentiment analysis is classified into three categories: document level, sentence level, and aspect level [3, 4]. Sentiment analysis done at document and sentence level does not represent the specific view of the reviewers instead provides generalized to entire product. Otherwise, the user has to be prompted manually for each specific aspect to review makes the process static. This work focuses on automation for aspect-level sentiment analysis.

There also exists another approach with supervised learning that focuses on opinion target extraction (OTE) and sentiment polarity towards a target or a category [8]. For OTE, a conditional random field model is proposed with several groups of features including syntactic and lexical features. For polarity detection, a logistic regression classifier is utilized along with the weighting schema for positive and negative labels, and several groups of features are extracted including lexical, syntactic, semantic, and sentiment lexicon features [5–7]. In the work conducted using bag-of-words [10, 13, 14], a semantic-based approach is used to list the word clusters, and supervised machine learning algorithms are used to generate the results.

3 Automated ASBA

Our goal is to generate sentiment rating for each aspect from the real-time reviews. The main challenges for the proposed system are limited resources, and the use of informal language in the reviews, that being the reason, proposed system using the available resources. The architecture of the proposed system is shown in Fig. 1, which contains five main tasks.

The architecture illustrates the general structure of the system and each component of the system which deals with each sub-task of the aspect-based sentiment analysis. This system is mainly developed using natural language processing and machine learning techniques in its core. The main aim of this work is to develop a dynamic approach to aspect-based sentiment analysis. The work is conducted exploring various techniques used in opinion mining and sentiment analysis. The

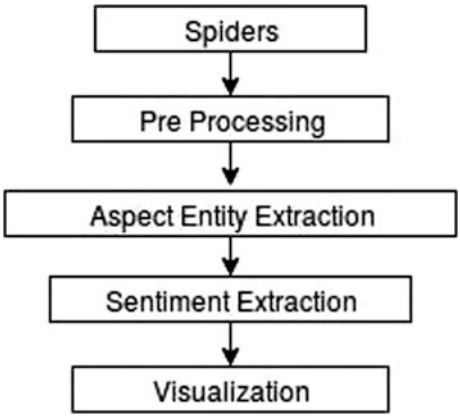

Fig. 1 Architecture of the proposed system

tasks in the architecture can be divided as four modules named data scrapping, pre-processing, aspect entity extraction, and sentiment extraction.

The data scraping module mainly concentrates on an automated process to crawl the web, to find the product details and reviews from multiple domains, and scrape them. The pre-processing module concentrates to prepare the data in an efficient manner to achieve high accuracy while extracting aspect entity pairs. Aspect entity extraction deals with mining features of the products as aspects and opinions expressed on them as entities. We used pattern mining techniques to identify and extract aspect entity pairs. Sentiment polarity is obtained by utilizing Naive Bayes classifier which is trained on a large movie review corpus.

3.1 Data Scraping

Web scraping is a technique used to retrieve data from the web using scripts called as crawlers or scrapers. In our work, we implemented scraping using Scrapy, a python package. The goal is to automate our system in order to scrape the data of same product in multiple domains (Amazon, Flipkart). We give real-time data from web sources as an input using this process. Web scraping begins with a target URL to the data source. The ids used in HTML tags will act as indices for the scraper. Those ids are used to extract the information required which can be text, links, images, etc. There are two formats, one uses CSS tags, and the other uses Xpath. The code snippet 'response.xpath('//title/text()').extract()' is used to scrape the data using Xpath.

3.2 *Pre-processing Reviews*

Pre-processing the scrapped reviews is of highest importance. The accuracy of the classifier majorly depends upon the preparation done in this stage. Real-time data is dirty, unstructured, and not properly formatted [12]. Given these challenges, we implemented the following techniques to get data cleaned.

(1) *Removing unnecessary characters*: Real-time reviews contain unnecessary characters such as exclamations (!), apostrophises (‘), underscores (_), hyphens (-), and Unicode characters. We used regular expressions to retrieve only text from a review. [a-zA-Z0-9_\’] characters which do not match the above regular expression will be removed from the review. Conventional NLP techniques such as stop word removal can be used, but comparatively our proposed model yields better outputs.
(2) *Stop word removal*: This is a technique in natural language processing where unnecessary words are removed from the text if they are listed in stop words, e.g., that, them, a, an, etc. This step drastically improves the efficiency of the system. The goal of this step is to narrow down the choice of words to get the focus words of the sentence, here aspects and their corresponding entities. We use stanford-NLTK package in python, stands for natural language toolkit. It provides a set of stop words and also provides the option to add additional stop words which can be domain specific.
(3) *Stemming*: For grammatical reasons, we use same root word differently at different occasions product, products–product [root word]

Example usage: ‘*I love this product, We loved using this product*’

In each of the aforementioned cases, product, products are the targets and love, loved, loving express the opinion on it. Instead on having two noun forms of the same word ‘product’, stemming will give the root word for it. By following this process, we can accumulate different noun forms under the same word, thus reducing the number of aspects and increasing accuracy.

4) *Parts of speech tagging*: This is most important technique to be followed when processing textual data. We use Stanford POS-tagger, CoreNLPTagger from the NLTK package. The functionality of the above-mentioned taggers is to tag a word in a sentence with its respective parts of speech. The following Fig. 2 depicts an example of parts of speech tagging.

For a well-structured sentence, parts of speech will be tagged perfectly. So, it is highly important to clean the reviews before tagging them. Stop word removal can be done even after the reviews are tagged. This helps us to preserve the POS tag of important words and get high accuracies while further processing. In the next step, we use pattern matching techniques to mine aspect entity pairs.

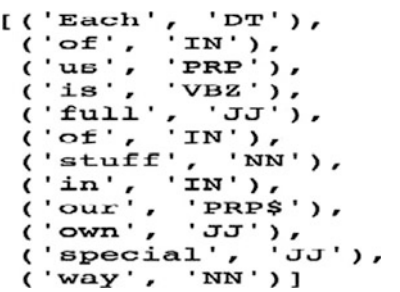

Fig. 2 An instance of POS tagging example

D. *Aspect Entity Extraction*

This section focuses on the core part of the system aspect entity pairs extraction. We used a pattern mining-based approach to achieve this. The POS tags 'NN', 'NNP' stand for noun and proper noun. We identified that aspects are nouns in the sentences and corresponding adjectives 'JJ', adverbs 'RBR/RB', verbs 'VBN' will act as entities to it. The following is an example:

Review: RAM is good, processor is amazing. Best laptop.

POS tagged: [('RAM', u'NNP'), ('is', u'VBZ'), ('good', u'JJ'), ('processor', u'JJ'), ('is', u'VBZ'), ('amazing', u'JJ'), ('Best', u'NNP'), ('laptop', u'NN')]

A-E pairs: ['RAM—good', ('NN','JJ')], ['processor—amazing', ('NNP','JJ')], ['laptop—best', ('NN','NNP')].

The system will first check if the pos tagged review contains any nouns, and then it checks for the corresponding entities associated with it. A word is considered as aspect if it has to be a noun, or it must not be a stop word, or it must appear more number of times. We observed that nouns which are repeated more number of times tend to become more important aspects. A counter is initialized, and appearance of aspects is counted. Similar aspects are grouped together along with their corresponding entities.

3.3 Sentiment Polarity Extraction

Sentiment of an aspect entity pair is predicted using TextBlob, a package developed on nltk and pattern libraries. This is achieved with a pre-trained sample on large dataset of movie reviews by Stanford. We used it to predict sentiments of smaller versions of reviews. Since the smaller versions of reviews only contain [Aspect + Entity], they are classified by textblob irrespective of the domain. For example, almost all the A-E pairs are in this form.

Example: ['camera','awesome'] ['battery','bad'] ['quality','great'].

The sentiment property returns a named tuple of the form sentiment (polarity, subjectivity). The polarity score is a float within the range [−1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective, and 1.0 is

very subjective. Here, we consider the polarity only. We display the result based on the polarity values, 0.0 is termed as neutral, values in the range [0.0, 1.0] are termed positive, and values between [−1.0, 0.0] are termed negative.

4 Results and Discussions

Previous works contributed to ABSA were performed on tagged data which is used to train and test using the models. The performance is measured accordingly with testing data. However, the work done on real-time techniques is low and is mostly confined to restaurants domain. To put it simply, our proposed system will extract sentiments of aspect entity pairs automatically without any definite class or category of the sentence.

We deployed our application in cloud to use it in real time. We used Amazon web services platform to deploy our application. There were numerous challenges to achieve this task. Some of them are package permissions, access permissions, cache management, etc. All these challenges were met with different strategies. Timestamps were used to keep the data intact in the system for a limited period. After the expiry, data (reviews of a product) are scraped again so that the system is updated with new reviews.

Given a product name as input, our system scraped the reviews, processed them, and generated results. The total time taken is about 14–15 s. To reduce the overall runtime, we implemented caching technique discussed earlier. We chose Amazon as the standard and scraped it first, and the spider utilizes the search results of Amazon. They contain the product name and top ten links for the product. The product name which is returned by Amazon spider is used to search Flipkart for the same reviews. This is how the system is maintaining its integrity in scraping data. Following are outputs of the system for the product 'iPhone 8'.

Figure 3 depicts the plot of positive and negative sentiments of top 15 aspect (*X*-axis) for combined reviews (Amazon and Flipkart combined). After extracting the aspect entity pairs, they are stored as (key, value) pairs. Key being the aspect and values is corresponding entities. A counter will be initialized to list the top n aspects. Then, each aspect is concatenated with its corresponding entities and is fed to the classifier to get the sentiments. The outputs vary from a range [−1.0, 1.0], −1.0 being absolute negative and 1.0 being absolute positive. Means of positive and negative sentiments of aspects are taken separately and displayed in the graph. This will help the users to better understand the positive and negative sentiments on each aspect of a product. The following figures Fig 4 are word clouds generated from all the aspects and their sentiments by the system. As discussed earlier, the focus of the system is to mine the aspects dynamically. On viewing the results, we can say that our system has mined potential aspects successfully. Following section contains evaluation of the results.

We use different dataset to evaluate our model. It is used in SemEval-16 task-5 [4]. The dataset consists of 3048 laptop reviews out of which 2373 reviews are

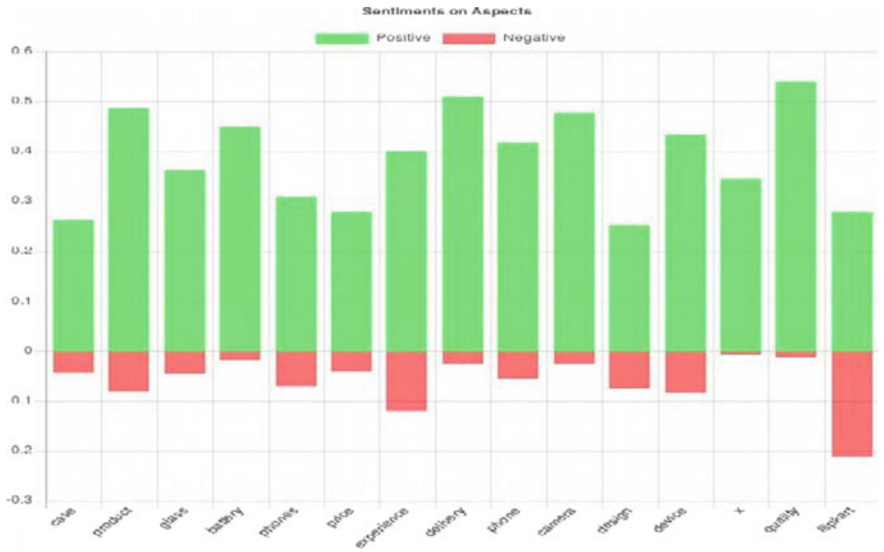

Fig. 3 Bar plot of aspects versus sentiments

Fig. 4 Positive and negative word cloud

tagged with aspects of each review. We gave the data as input to our system and recorded the results. Interestingly, our system listed out new aspects which were not listed in the dataset earlier. Some of these new aspects include HP, Macbook, OS, laptop, etc. We processed the dataset and listed out top 50 aspects and compared the results of top 50 dynamic aspects of our system. We evaluated the results based on precision, recall, and F1-measure for aspect extraction.

Following table will list out the results. Our model is termed as ABSA-dy which is compared with the best results from the work done by Nipuna and Chamira [2]. The analysis [2] performed on laptop review dataset is done by using supervised learning approaches including unigrams, stemming, pos tagging, and Headword POS. The best results observed are mentioned in the above table with the

Static	Dynamic
screen	computer
use	laptop
battery life	screen
programs	windows
price	keyboard
keyboard	mac
battery	warranty
software	software
features	thing
hard drive	battery
size	problem
performance	pc
speed	drive
warranty	price
memory	hours
quality	macbook
applications	programs
vista	netbook
Windows	quality
graphics	machine

Model	Domain	Recall	Precision	F-Measure
ABSA-dy	laptop	0.862	0.735	0.793
ABSA-su	laptop	0.508	0.451	0.478

TABLE I: Flipkart and amazon data

Fig. 5 Comparing static and dynamic aspects

label ABSA-su. We observed a great enhancement in the measures using the dynamic approach to mine aspects. Following Fig. 5 will show the comparison between static (mentioned in the dataset) and dynamic aspects (generated from our system).

We compared the results of our system for the product 'iPhone 8' in three modes: using only Amazon reviews, only Flipkart reviews, and using both the reviews. We observed that there are slight differences in positive and negative sentiments expressed. Dynamic aspects generation can be used by various vendors to support their marketing strategies and customers to better understand the products.

5 Conclusion

In this paper, we present a composite method based on pattern mining combined with natural language processing to solve the ABSA problem in real time. To improve the accuracy, we implemented a technique to append new stop words to the system. We were able to bring down the running time to mere 12 s including the tasks of scrapping and processing from two domains (web sites). Dynamic aspect classification can be viewed as a better technique than static and supervised techniques as it is efficient with an F1-measure of 0.793. The system obtained very satisfying results for dynamic aspect extraction from real-time reviews on multiple domains.

References

1. Gojali, S., Khodra, M.L.: Aspect based sentiment analysis for review rating prediction. In: Proceedings of International Conference on Advanced Informatics: Concepts, Theory And Application, IEEE, Malaysia (2016)
2. Pannala, N.U., Nawarathna, C.P.: Supervised learning based approach to aspect based sentiment analysis. In: Proceedings of IEEE International Conference on Computer and Information Technology (CIT), Fizi (2016)
3. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: Aspect based sentiment analysis. In: Proceedings of Proceedings of the 9th International Workshop on Semantic Evaluation SemEval 2015, Denver, Colorado (2015)
4. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I.: Aspect Based Sentiment Analysis. SemEval (2016)
5. García-Pablos, A., Cuadros, M., Rigau, G.: W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst. Appl. **91**, 127–137 (2018)
6. Ruder, S., et al.: INSIGHT-1 at SemEval-2016 Task 5: deep learning for multilingual aspect-based sentiment analysis. SemEval@NAACL-HLT (2016)
7. Jebbara, S., Cimiano, P.: Aspect-based sentiment analysis using a two-step neural network architecture. In: Cognitive Interaction Technology Center of Excellence (2016)
8. Hamdan, H.: Opinion target extraction and sentiment polarity detection. In: Oroceedings of SentiSys at SemEval-2016 Task 5 (2016)
9. Xu, L., Lin, J., Wang, L., Yin, C., Wang, J.: Deep convolutional neural network based approach for aspect based sentiment analysis. Adv. Sci. Technol (2017)
10. Guha, S., Joshi, A., Varma, V.: SIEL: aspect based sentiment analysis in reviews. In: 4th Joint Conference on Lexical and Computational Semantics (2015)
11. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report. Stanford **009**, 112 (2017)
12. Pradhan, V.M., Vala, J., Balani, P.: A survey on sentiment analysis algorithms for opinion mining. Int. J. Comput. Appl. (2016)
13. Zhang, W., Xu, H., Wan, W.: Weakness finder: find product weakness from chinese reviews by using aspect based sentiment analysis. Expert Syst. Appl. **39**(11), 10283–10291 (2012)
14. SoufianJebbara, P.C.: Aspect-based sentiment analysis using a two-step neural network architechture. Bielefeld University, Germany (2016)