FEASA: AN APPROACH TO RESOLVE SENTIMENT AND GENERATE FEATURE EXTRACTION MATRIX

AmeenaShad*, Manjunath H R
*PG Student M.Tech Computer Science, ShriDevi Institute of Engineering and Technology Tumakuru, Karnataka, India
Assistant Professor Computer Science, ShriDevi Institute of Engineering and Technology Tumakuru, Karnataka, India

DOI: 10.5281/zenodo.50840

KEYWORDS: Sentiment analysis; feature extraction matrix; reviews; FEASA (Feature extraction and Sentiment analysis).

ABSTRACT

Sentiment analysis is the study of classifying human's sentiments, evaluations, attitudes, opinions about some topic, product, expressed in form of text or speech. As e-commerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. In order to improve the customer satisfaction, many e-commerce sites provides the provision to write reviews about products. Instead of manually reading and evaluating numerous reviews, an automated procedure can be helpful and can be easy to get the overall polarity of the reviews for the product. The goal of the research is to present an approach with the target of deriving qualitative sentiment analysis which will be helpful for recommendation. This paper introduces the FEASA (Feature extraction and sentiment analysis), an approach to resolve sentiment and generate feature extraction matrix for each product. FEM will help in determining the features, being commented upon and also help in detecting the popular features among the reviews. The paper aims at mining sentiment for each product review and summarizes overall sentiment score associated with each product review. The efficient ranking of the features will be deduced, based on the opinions.

INTRODUCTION

Opinion mining is another term used for sentiment analysis. It is the study of analyzing one's sentiments, evaluations, attitudes about any topic. Before the year 2000, there was little research done in the field. But now there are several reasons that is the reason why the field has become a hottest research area. Its applications can be seen in every domain. It is important for both individuals as well as for the organizations to grow, hereby providing the strong motivation for research. We have large volume of data available on web. The inception of the sentiment analysis coincides with those of online data in the form of reviews, blogs. Without the availability of that data, the research on sentiment analysis could not have been possible. Here, the focus is, not only to detect the polarity of the product reviews but to resolve the sentiment at more detailed level. The intended work is to extract the features from the review and detecting the polarity for each aspect, thus resulting in feature extraction matrix (FEM). The work is also concerned with calculating the sentiment score associated with features of entity. Here, a FEASE (feature extraction and sentiment analysis) has been proposed to achieve the objectives. A subjective classifier is used to separate the subjective sentences. Subjective sentences contain the sentiment bearing words. By representing the feature extraction matrix, it would be easier for other readers to understand on what features the opinion holder has commented upon.

RELATED WORK

Sentiment classification comes under the problem of text classification. Previously, text classification incorporated a work of classifying a document topic wise, e.g., sports, and sciences. Key features help in detecting the theme of the document. As far as sentiment analysis is concerned, sentiment words or opinion words, for example good, excellent, amazing, bad etc, play a significant role in classifying a document. These are the words that help in deciding the polarity of the reviews. In [1], the problem of entity discovery and entity assignment was studied out. The researchers proposed the techniques for entity discovering and entity assignment in the sentences. Pattern-based methods were employed to solve the entity discovery problem. Also, the comparative sentences were studied for the assignment of the entities. The unsupervised approach was presented for classifying reviews in [2]. Average semantic orientation of the phrases was used for classifying the reviews. Mining the various product features and customer reviews summarization was studied in [3]. The task also involves the finding of the opinion sentences and deciding the polarity of the sentences. The main task was to provide the brief summary of the customer reviews. In [4], the identification of opinion holder and sentiment about any topic was studied. Word sentiment classifier was employed to determine the sentiment. In [5], a holistic lexicon-based approach was proposed to opinion mining. The research focused upon context dependent analysis of the opinions. In [6], the novel rule mining and supervised learning approach was presented to identify the comparative sentences. Three types of documents were taken into analysis namely, news articles, forums, and reviews. For web opinion mining and extraction, a machine learning approach was implemented in [7]. Multiple linguistic features (e.g., part-of-speech, phrases', and surrounding contextual clues of words/phrases) were employed to perform analysis on web opinion. In [8], the sentiment analysis carried out on subjective sentences, discarding rest of the sentences. Subjective sentences are the sentences that contain opinions. The minimum cut approach was implemented to design the algorithm for analysis on movie reviews. In [9], a method based on bootstrapping was employed for studying targets and opinions on them. The method was called double propagation method. The method aimed at determining the linking between the opinion words and targets. Various machine learning methods were studied in [10]. The methods were naïve bayes, maximum entropy model, and support vector machines. The movie-review domain was selected for analysis and experimentation of these methods. Support vector machines tend to be the best method in the result. In [11], the detection of pro and con reasons of reviews is studied. The aim was to analyse why people like or dislike that product. The maximum entropy model was used to perform this task. In [12], the genetic algorithm was proposed to classify the documents of different languages into desired sentiment. Sentiment classification methodologies were applied on two languages namely, English and Arabic web forums. For efficient feature selection, the entropy weighted genetic algorithm (EWGA) was employed. In [13], a semi-supervised approach was used to sentiment classification. The target here was to mine the unambiguous reviews using spectral techniques and then ambiguous reviews using novel combination of active learning, transductive learning, and ensemble learning. In [14], the customer feedback data was analysed for sentiment classification. In [15], the different weighting schemes are studied for sentiment classification. The proposed work is to calculate the overall sentiment score associated with each product review and also score related to each aspect of the entity.

PROPOSED SYSTEM

In the proposed approach, the algorithm will first get the reviews of products from the given URL and then parse the reviews to clean them. Find the positive and negative polarity for each review against the product. The product is again rated on the various attributes namely Screen, Phone, Price, Speaker, Battery, Camera and Quality and then provides the overall sentiment distribution of product. With this paper, the aim is to further work on this automation process using a combination of data aggregation techniques, NLP, linguistic analysis and popular visualization techniques we generate visually appealing and easy to understand graphs which provide summarized feedback. This is done by performing detailed sentiment analysis on the data.

The system performs the task in three steps: (1) the opinion sentences are identified and then decided whether opinion sentence is positive or negative; (2) the product aspects are then identified that have been commented by the opinion holders; (3) feature extraction matrix (FEM) is produced.

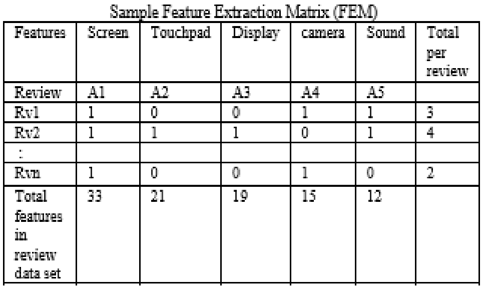
The sub-steps of the algorithm are discussed below:

Opinion Extraction and Opinion Orientation Identification: The system crawls the product reviews and put them into the review database. The opinion words are extracted from the review and semantic orientation of the words are predicted. The algorithm identifies the opinion bearing words. The opinion bearing words are the words which expresses opinion about anything. It has been studied that adjectives play a vital role in expressing opinions. Presence of adjectives defines the subjectivity in the sentences. The system then aims at finding the semantic orientation of the found opinion words which will be then used to determine the polarity of the review. The algorithm classifies the review as recommended if positive things are said about the product in the review and would classify it as not recommended if negative things are being said about the product. A subjective classifier is used to separate the subjective sentences. Subjective sentences contain the sentiment bearing words. To calculate the numeric score for the sentiment bearing words, a lexicon is being used. It contains around million of words. Whenever the review is consider for counting score, our algorithm will undergo the process, where sentiment bearing words would compared with the words listed in the lexicon. The resultant is the numeric score on the basis of which, a product can be resolved to the desired orientation. Furthermore, when a customer talks about various aspects of the product, it is then become necessary to calculate the sentiment score related to each feature.

Feature Extraction and Feature Orientation Identification: The algorithm aims at identifying and extract object features that have been commented on by an opinion holder, followed by feature pruning task. For this purpose, we have made our own lexicon of features from which features would be extracted and a desired sentiment for that particular aspect is calculated. It would determine whether the opinions on the features are positive or negative. Based on the opinions expressed with each feature that has been commented upon by the opinion holder, a sentiment score get associated with the feature by making use of the lexicon approach.

Generation of FEM (Feature Extraction Matrix): Going further, a feature extraction matrix will be generated which would tell the number of features of the product are present in the review and the polarity score of the features present in the review. The sample feature extraction matrix (FEM) is depicted in Table 1. In case, the customer talks about the new feature, which is not present in the feature set, the new feature is added in the feature set. The information obtained from FEM can be helpful in deducing efficient ranking among the aspects which have been commented by the opinion holder. The popular aspects can also be mined using the information. Final step would show the graphical representation of overall sentiment about the features and the product.

Table I
Sample Feature Extraction Matrix (FEM)



Features	Screen	Touchpad	Display	camera	Sound	Total per review
Review	A1	A2	A3	A4	A5	
Rv1	1	0	0	1	1	3
Rv2	1	1	1	0	1	4
:						
Rvn	1	0	0	1	0	2
Total features in review data set	33	21	19	15	12	

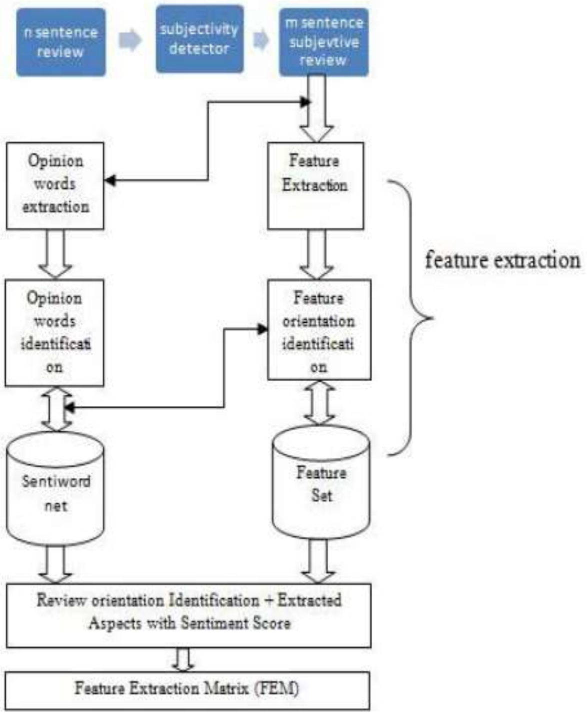

n sentence review
subjectivity detector
m sentence subjevtive review
Opinion words extraction
Feature Extraction
feature extraction
Opinion words identificati on
Feature orientation identificati on
Sentiword net
Feature Set
Review orientation Identification + Extracted Aspects with Sentiment Score
Feature Extraction Matrix (FEM)

Fig 1: The FEASA

SYSTEM DESIGN

The FEASA is a feature extraction and sentimental analysis system. A review about a product is fed as the input and the Feature Extraction Matrix (FEM) generation algorithm. FEM matrix has each row as an observation for a product and each of the columns represent the feature and also generates a Ranking, which is used for ranking the products based on search criteria matching the FEM matrix.

Data Preparation:
The data preparation involves the data pre-processing and cleaning of the data set where the Collected Reviews will be considered as input and Output will be Cleaned reviews which does not contain any not meaningful words. Pre-processing steps include removing of information about the reviews that are not required for performing sentiment analysis, such as date, time of a review.

Review Analysis:
The review analysis step is used to identify the necessary information including opinions and product features. The input will be cleaned reviews which thereby generate an output of computation of frequency across all products and all reviews per feature. Opinions and features are extracted from this step.

Sentiment Classification:
The last step is to detect the polarity of the document. Technique is used to assign the sentiment to the opinions given by the people.

Algorithm FEASA ()
Input: Review data set
Output: Each review orientation identification and extracted aspects with associated sentiment score and FEM

A. Procedure Opinion Words Extraction () and Opinion Orientation Identification ()
1. Let R be the set of reviews i.e. R = {r1, r2, r3....... rn}, where each review r_i€ R .

```
2. Spilt each review $r_i$ into a sentence set S. S is composed of number of sentences i.e. S = {s1, s2, s3.... sm}. Each sentence $s_j$€ S.
3. For each sentence $s_j$€ S do,
            extract opinion bearing words
                 for each word $w_q$ in the sentence do
                          traverse the lexicon to get the numeric polarity score
                          If (score = exist)
                          SO ($w_q$) = recorded
                          // SO i.e. semantic orientation of each opinion bearing word is recorded
                         end if
            Sentence Score ($s_j$)  =$\sum^{p}_{q=1}$ SO($w_q$)
                  end for
4. Total Review Score ($r_i$) =$\sum^{m}_{j=1}$Sentence Score($s_j$)
        // TRS (Total Review Score) now contains the polarity numeric score of the review.
```

B. Procedure Product Aspects Extraction () and Identification ()

```
1. Features are searched against the self constructed feature set.
2. The sentiment score of the extracted features are dependent on the opinion words that are associated with them in a minimum distance.
3. Once the polarity score for the words are identified. The next is to associate them with the relevant features.
          for each feature $f_x$ belonging to sentence $s_j$,
                 traverse for the opinion words $w_q$  which are more closer to it.
                          $F_x$ = SO ($s_j$)
          end for
```

C. Procedure Generating Feature Extraction matrix (FEM)

```
1. In the matrix M[r][f], consider review set R as tuples and feature set as columns.
2. for all review $r_i$€R, do
         If (SO ($f_x$) = exists)
                 Set value (M[$r_i$,$f_x$]) = 1
         else
                  Set value (M[$r_i$,$f_x$]) = 0
    end for
```

CONCLUSION

In this paper, we proposed an approach for resolving sentiment at aspect level. It is necessary to identify the sentiment related to each aspect of entity, when review discusses about several aspects of the entity. There is a need to calculate the overall sentiment of the product, more accurately. A feature extraction matrix (FEM) would be generated as the result of the proposed work to determine sentiment related to features of the product. The work also includes the task of comparing the products on the basis of the product reviews and finding out the popular aspects among the reviews. As size of information present on the internet has taken a shape of the giant it has become a necessity to increase the efficiency of the search engines. Web mining is aiming in this direction. In this paper, we have done both feature based on negative, neutral and positive polarity.

REFERENCES

1. X. Ding, B. Liu, and L. Zhang, "Entity discovery and assignment for opinion mining applications," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09),* 2009.
2. P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the Association for Computational Linguistics*, pp. 417-424, July 2002.
3. M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), Aug. 2004.

4. S. M. Kim and E. Hovy, "Determining the sentiment of opinions," in Proceedings of Interntional Conference on Computational Linguistics (COLING'04), 2004.
5. X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the Conference on Web Search and Web Data Mining (WSDM'08),* 2008.
6. N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'06). 2006a.
7. W. Jin, H. H. Ho, and R. K. Srihari, "OpinionMiner: a novel machine learning system for web opinion mining and extraction," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discoveryand Data Mining (KDD'09),* 2009b.
8. B. Pang and L. Lillian, "A sentimental Education: Sentiment analysis using Subjectivity Summarization based on minimum cuts," in *Proceedings of the Association for Computational Linguistics,* pp. 271-278, 2004.
9. G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," *Computational Linguistics,* vol. 37, No. 1: 9.27, 2011.
10. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP'02),* 2002.
11. S. M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in *Proceedings of COLING/ACL 2006 Main Conference Poster Sessions (ACL'06),* 2006.
12. A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems (TOIS),* vol. 26, Jun. 2008.
13. S. Dasgupta and V. Ng, "Mine the easy, classify the hard: a semisupervised approach to automatic sentiment classification," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP(ACL'09)*, pp. 701-709, Aug. 2009.
14. M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of International Conference on Computational Linguistics(COLING'04), 2004.*
15. J. Kim, J. J. Li, and J. H. Lee, "Discovering the discriminative views: Measuring term weights for sentiment analysis," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL'09)*, pp. 253–261, Aug. 2009.