International Journal of Research in Engineering, Science and Management
Volume-3, Issue-2, February-2020
www.ijresm.com | ISSN (Online): 2581-5792

Feature Specific Sentiment Analysis on Product Reviews

A. Austin Solomon Raj[1], S. K. Sudhakar[2], M. Nandhini[3], R. P. Vijai Ganesh[4]

[1,2,3]*Student, Department of Computer Science Engineering and Technology, Dr. Mahalingam College of Engineering and Technology, Coimbatore, India*

[4]*Assistant Professor, Department of Computer Science Engineering and Technology, Dr. Mahalingam College of Engineering and Technology, Coimbatore, India*

***Abstract*: Most of the users critically review anything on the internet to show case their humble opinion. These opinions contain more valuable information which helps in decision making. People who purchased products express their opinions concerning their satisfactions and criticisms. Due to enormous amount of web reviews available for a product, it is extremely time-consuming and difficult to manually analyze the review and come to a conclusive decision. Reviews generally contains product feature specific factual information along with the opinion sentences which may be positive or negative. The text reviews obtained from most of the web review sources found to be unstructured and necessitates the automatic identification of the opinion and also the identification of the explicitly visible and implicitly present product feature associated with the opinion sentence. Analyzing and extracting the actual opinion throughout the reviews manually is very difficult. So, an automated methodology is needed to solve this problem. The Aspect based opinion mining is such a methodology which describes the important aspects of each opinion and classify them on their polarity.**

***Keywords*: SVM, Aspect based opinion classification, Summarization, Polarity prediction, Review classification, Rating prediction.**

1. Introduction

Mining valuable information from these product reviews not only provides some necessary purchase information for the potential consumers but also helps producers track the feedbacks of users on time. The feedback information contributes producers to maintain the good characteristics of products and improve the inferior products timely, and finally make them gain competitiveness in the near future. However, the huge number of network comments also makes mining useful information to be a new challenge. It is hard and unrealistic for human to tackle all the reviews and classify them to positive or negative manually. Under this situation, automatic sentiment analysis technology has significant meaning. Via this approach, we can automatically extract the opinion which the author expressed on the product or its features. Document-level and sentence-level [2] based sentiment analysis are usually used to judge the overall evaluations of the special product. In order to obtain much finer grained and detailed product information, feature-level sentiment analysis is essential [3]. There are two major topics involved in feature-level sentiment analysis [4]:

1. Extracting the product features which the user concerned. For example, in the sentence “The camera quality of this mobile phone is good”, (tone quality) is the product’s feature.
2. Analyzing the sentiment orientation of the product features. That is, classifying the author’s evaluations expressed on the product feature into positive emotion, negative emotion or neutral. In the above sentence, the evaluation on the feature “camera” is positive.

2. Literature Survey

A. *Opinion Mining Approaches*

Sentiment can be extracted through either by Machine learning approach or by Lexicon-based approach or Semantic approach or Statistical approach or Rule based approach [2]. Machine learning techniques can be categorized [4] as supervised and unsupervised learning whereas lexicon-based approach techniques can be categorized into corpus-based approach and dictionary-based approach. Support Vector Machine (SVM) outperform Naïve Bayes classifier.

B. *Aspect Based Opinion Mining*

Aspect based opinion mining is preferred in this work. The core tasks in aspect based opinion mining is aspect identification, aspect based opinion word identification and its orientation detection.

In [1] SentiWordNet, two word phrases and linguistic rules together for opinion orientation detection, with automatic acquisition of aspects. In this work only explicit aspects are considered and word sense disambiguation is ignored.

In [3] Deep learning approach to improve aspect based Mining is introduced. CNN is comprised of one or more convolutional layers which are responsible for major breakthroughs in image classification. More recently, CNN is also applied to problems in Natural Language Processing like information retrieval and relation classification, sentiment analysis spam detection or topic categorization. Sentences or documents that are the input of most NLP tasks can be represented as a matrix where each row represents one token. A token may be a word or a character. The convolutional layers can be represented as the weighted sum of the word vectors with respect to the shared weight matrix.

The dependency parsing [1], which reveals the syntactic structure by analyzing the dependency relationship between different components of a sentence. That is, the dependency parsing labels grammatical constituents such as "subject-verb", "verb-object" and analyzes the relations between them.

Topic models [5] are most popularly used aspect extraction methods as, web documents have reviews with a mixture of topics whereas each topic is a probability distribution of words. Topic models are adapted and extended by probabilistic Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) models. Topic modelling [6] or clustering is only able to find some general or rough features, it is difficult to extract finer-grained or precise features.

In [7] Sentiscore algorithm is proposed. The classification process is done by KNN algorithm. 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 better accuracy but complexity increases.

C. *Summarization and Rating Prediction*

In [8] five steps to generate the summarization is discussed: 1) obtain the review text; 2) extracts adjective-noun word pairs; 3) counts each word pair's occurrences; 4) performs a sentiment analysis of each word pair; and 5) displays the word pairs. TF and TF-IDF scoring methods are introduced but tend to contain word pairs which users would find rather irrelevant or uninformative.

Unlike previous works that used only Multinomial Naïve Bayes model This work [9] employed Bigram and Bigram-Trigram Multinomial model. BigramTrigram Multinomial model reported on the par best accuracy given by Random Forest while showing 12 times faster performance. Thus, Bigram-Trigram model is efficient model in text classification when dealing with huge dataset.

3. Proposed Methodology

In the proposed model, we use the machine learning algorithm (linear SVC) to classify the feature of the product. In which each feature of the products are trained, which makes the model more accurate and specific for the product. In preprocessing of data, the multiclass sentence is separated to improve the efficiency of the model

In addition, Opinion summarization of each individual feature is done by the dependency parser. The rating for the review are computed by the random forest algorithm and the feature for the model is extracted by the bag of word.

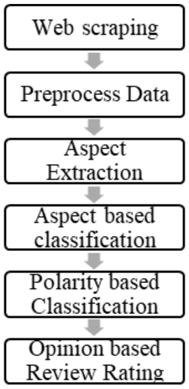

Web scraping
Preprocess Data
Aspect Extraction
Aspect based classification
Polarity based Classification
Opinion based Review Rating

Fig. 1. Proposed method

A. *Web Scraping*

Web Scraping (also termed Screen Scraping, Web Data Extraction Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

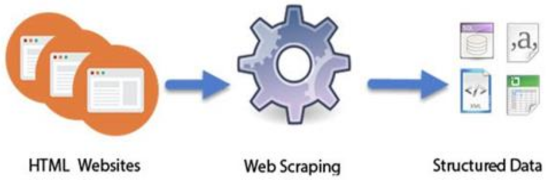


Fig. 2. Web scraping

B. *Preprocess Scrapped Data*

Most often sentiment analysis is fed from natural language sentences. While generating the content using natural language, there may be incomplete texts and malformed sentences are included. Therefore, in order to do sentiment analysis, the sentences in the natural form, need to be pre-processed.

Pre-processing the review data is the exercise of cleaning and removing the noise from the sentences. Since online texts are customer feedback, there are lots of noise such as smiles and emoticons. So, keeping those unnecessary words in the review database makes more difficult, as each word in the text is treated as an input to the system. So, it is important to reduce the noise in the text. As a result, the performance of the classifier and the speed of the classification process will be increased. Reviews from the web contains more noisy data like '\n' and non-opinion sentences. These sentence are identified and removed by the regex parser in the NLTK. The Multiclass(features) sentence are separated by finding the connecting sentence (CC) and the DT using the pos-tag to increase the performance of the model.

C. *Aspect Extraction*

Aspects are the important features mentioned by the customers in the online review. They give their feedback based on these aspects. However, these features are hidden within the sentences. Generally, customers do not clearly mention that a sentence is based on a particular aspect in the online review.

Approaches in Aspect Extraction are 1. Frequency based approach, 2. Relation based approach 3. Model based approach. Here the Frequency based approach and Relation based approach are used to identify the feature. The frequency based approach is used to find the features needed for the machine learning algorithm to predict the labels and the Relation based approach is used for dependency parser.

1) Frequency based Approach

This approach identifies the frequent aspects of product on which many people have expressed their opinion. If, the occurrence of aspect related terms is more than that of the pre-defined threshold value, then that term is considered to be frequent aspect. In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

TFIDF score for term I in document j=TF(i,j)*IDF(i) where
IDF=Inverse Document Frequency
TF=Term Frequency

$$TF(i,j) = \frac{\text{Term I frequency in document j}}{\text{Total words in document j}}$$

$$IDf(i) = \frac{\text{Total Documents}}{\text{documents with term i}}$$

and
t=term
j=document

2) Relation based approach

In this domain, aspects are the important features of the mobile. Aspect can be a word or a phrase. For example, “battery life”, “camera quality”, “design” are the main aspects of the mobile domain. Using a POS tagger, preprocessed review files were tokenized. For that, POS tagger was given to each word of the review file.

For example, the review “very good camera” is tokenized as “very/jj good/jj camera/nn” where jj and nn represents adjective and noun respectively. When author reads reviews manually in most situations, nouns and noun phrases are identified as aspects. In order to extract aspects from the sentences in all reviews, observing for nouns and noun phrases are needed. After tokenizing the review texts using a POS tagger, it extracts NN (noun, singular), NNS (noun, plural), NNP (proper noun, singular) and NNPS (proper noun, plural) tagged words from the review file as aspects dependency parser is used to extract the Aspect from the sentence and their corresponding opinion word by using the synaptic dependency. Thus summarize the feature of the product.

3) Dependency Parser

Opinion words express the opinion towards the aspects. So, during this phase, aspect related opinion words should be identified. Opinion words can be used as adjectives, adverbs and verb combinations. For the opinion word extraction process the dependency parser is used. Spacy provide a representation of grammatical relations between words in a sentence. These dependencies are triplets: Name of the relation, governor and dependent. For example, consider the review; “The camera quality is impressive and I love it”. According to the Spacy, the parsed output for this review is as follows:

```
the det quality NOUN []
camera compound quality NOUN []
quality nsubj is AUX [the, camera]
is ROOT is AUX [quality, impressive, and, love]
impressive acomp is AUX []
and cc is AUX []
i nsubj love VERB []
love conj is AUX [i, it]
it dobj love VERB []
```

Here the camera quality and its opinion impressive is indirectly connected. The children of the root connects the quality with impressive (The impressive is the acomp and the quality is the nsubj which is in compound with the camera). Thus the summarized text is “camera Quality - impressive”.

Table 1
Review Summarisation

Review	Summarization
“I bought the phone yesterday the camera is good and the battery last long”	Camera – good, Battery -last long
“suddenly touch display has been crashed ... it does n't working properly while swiping”	Touch display - crashed

D. *Aspect and polarity based opinion classification*

The Preprocessed data enters the trained Model which separates the opinion Based on the Aspect of the product for example if we process the Mobile phone data the aspects may be ‘Camera’, ‘Performance’, ‘Display’, ‘Battery’ and each aspects carries their own polarity whether the opinions on the aspects are positive or negative (‘+’, ‘-’) SVM is used to train the model which gives better accuracy than other classification algorithm.

Table 2
Aspect and Polarity Prediction

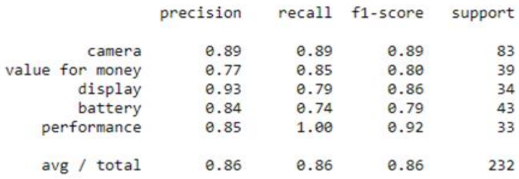
Reviews	**feature**	**Polarity**
Camera is good	Camera	Pos
Display is bad	Display	Neg
Performance is good	Performance	pos
Good mobile in the range	Value for money	pos
Camera quality is poor	Camera	neg
Display is good	Display	pos

E. *Opinion based product feature rating*

The Aggregate score for each product features are computed which gives an overall view of the product feature opinion in the review. The Bag of words is use to extract the opinion features from the review sentence and the Random Forest algorithm is used to predict the output result rating of the review. The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of words, disregarding grammar and even word order but keeping multiplicity.

4. Evaluation and Result

In this experiment we have used the natural language processing to separate the multiclass sentence into the single class sentence which increases the accuracy and the efficiency of the SVM model.



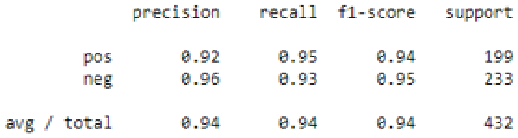
```
                 precision    recall  f1-score   support

        camera       0.89      0.89      0.89        83
value for money       0.77      0.85      0.80        39
       display       0.93      0.79      0.86        34
       battery       0.84      0.74      0.79        43
   performance       0.85      1.00      0.92        33

   avg / total       0.86      0.86      0.86       232
```

Fig. 3. Result of aspect prediction

The result of the Aspect classification is shown in fig. 3. The main Aspects of the mobile data are the camera, value for money, display, battery, performance. The SVM is used to predict the aspects of the mobiles which shows an accuracy of 86%.



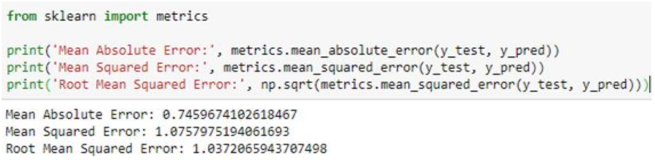
```
             precision    recall  f1-score   support

        pos       0.92      0.95      0.94       199
        neg       0.96      0.93      0.95       233

avg / total       0.94      0.94      0.94       432
```

Fig. 4. Result of polarity prediction

In fig. 4 the results of the polarity prediction are tabulated. The SVM is used to predict the polarity of the reviews. It shows 94% of accuracy.



```
from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 0.7459674102618467
Mean Squared Error: 1.0757975194061693
Root Mean Squared Error: 1.0372065943707498
```

Fig. 5. Result of rating prediction

Fig. 5 shows the error rate of the prediction is minimum. The random forest shows low level of error.

5. Conclusion and future work

From the above results we conclude that the frequency based aspect extraction is used to extracts the feature from the reviews and SVM classifies the aspects of the review and their polarity of the opinion. The SVM classifier gives an accuracy of 86% when the multiclass sentence is separated during the preprocess state which is higher than other machine learning classification algorithm like Naïve Bayes and Decision tree algorithm. The synaptic dependency between the words of the review are found and the aspects and their relative opinions are being extracted using the dependency parser. The feature specific sentiment analysis on the product reviews is achieved by the aspect based opinion mining method. The SVM classifier predicts the polarity of the reviews with an accuracy rate of 94%. The Rating score predictor for each reviews are done by the random forest model which gives a minimum error rate.

References

[1] I. K. C. U.Peera, H. A. Caldera , “On the level of various Aspects Based Opinion Mining on Restaurant Reviews using Automated Methodology”, IEEE International Conference on Computational Intelligence and Applications 2nd 2017

[2] Avinash Golande, Reeta Kamble, Sandhya Waghere, “On the idea of Feature Based Opinion Mining and its Limitations”, Springer International conference, September (2016).

[3] Paramitha Ray, Amlan Chakrabarthi, “On the idea of Rule Based method to Improve Aspect Level Sentimental Analysis and Mixed Approach of Deep Learning Method, 2019.

[4] Penubaka Balaji, D. Haritha and O. Nagaraja, “On the basis of Analysing Several Opinion Mining Techniques and Sentimental analysis”, International Journal of pure and Applied Mathematics, February 2018.

[5] Krishna B Vamshi, Ajeet Kumar Pandey, Kumar A. P. Siva, “On the Topic Model Based Opinion Mining and Sentimental Analysis”, IEEE International Conference on Computer Communication and Informatics (Jan. 4 2018, Coimbatore, India.

[6] Haiping Zhang, Zhengang Yu, Ming Xu, Yueling Shi, “On the rule based models Feature Level Sentiment Analysis for Chinese Product Reviews,” IEEE International Conference on computer Research and Development, 135-140, 2019.

[7] A. Nandhini, G. Vaitheswaran, L. Arockian, “On the A Hybrid approach for aspect based sentiment analysis on big data,” 2019.

[8] Koji Yatani, Michael Novati, Andrew Trusty, Khan N. Truong, “On the Analysis of Adjective Noun Word Pair Extraction Methods for Online Review Summarization,” 2012.

[9] Reddy, C. S. C., Kumar, K. U., Keshav, J. D., Prasad, B. R., and Agarwal, “Prediction of star ratings from online reviews,” TENCON 2017 – 2017, IEEE Region 10 Conference, 2017.