Aspect and Opinion Extraction from Unstructured Data using Machine Learning Techniques

Zaheeruddin Ahmed [1], Dr.B.L.Raju, Professor & Principal [2], Dr.Jatin Sharma, Charan Singh University [3]

Abstract: Now-a-days, online reviews in the e-commerce website are increasingly written by the consumers of the product. These reviews have become an important source of information for the new customers to research about these products online. The curious customer research often leads to decision making towards purchasing the product based on online reviews. These reviews reflect the experience of the user with the product over a considerable amount of time. Most of these data is saves in websites unstructured data. Retrieval and extraction of the information is essential works and importance in semantic web areas. Many of these requirements will be depend on the storage efficiency and unstructured data analysis . Merrill Lynch recently estimated that more than 80% of all potentially useful business information is unstructured data. We analyze the unstructured data individually and converted it into structured data collectively. Text mining and natural language processing are two techniques with their methods for knowledge discovery form textual context in documents. In this study, text mining and natural language techniques will be illustrated. The emerging field of opinion mining was investigated by Natural Language Processing (NLP) community for nearly two decades. In this work is inclined to feature level opinion mining of online reviews, in which the main purpose is to identify and extract product features and opinions. **Method:** The step-by-step feature extraction approach is followed to reach the goal of extracting maximum number of product features from the product reviews. Various types of nouns are extracted in the form of product features. These are namely frequent features, relevant features, implicit features and infrequent features. **Findings**: The results show that the comprehensive feature extraction approach performs better than the particular way for extracting the product features in the semantic environment. **Applications:** This approach is used in e-commerce websites to find out what product features are of interest to the customers. This model is useful in recommending products to the customers as the search for a product in the e-commerce site takes place, the features from the product reviews are helpful with the corresponding opinion orientations. This forms the basis for suggesting similar products using the calculated sentiments in the recommendation process.

Keywords: *Comprehensive Product Features Extraction, E-Commerce, Natural Language Rules, Online Reviews, Opinion Targets, Product Features, typed dependencies; contextual clues; opinion word sense*

1 Introduction

The information age is also known as digital, in which technology provides the ability to transfer information quickly. Current generations are seamlessly using web connectivity to share their opinions with others. The traditional information sharing is to exchange data between a sender and receiver. Younger generations are showing lot of interest to share the content in web, so this trend is growing faster and may increase in future. To share their preferences and opinions with others, most of the people are depending on web and are using blogs and comments to rate the product. The comments in the web blogs represent certain opinion of the person. This kind of data is useful to rate the product.

The user contribution is the prime value driver in most of the Web 2.01 social networking applications. This had a remarkable impression on the way the users interact with the web. The social commerce sites like Amazon allow the people to interact with the consumer written online reviews. These user contributed reviews express their opinions which are useful to the information readers. The language patterns that are inherent in the product reviews provide the crucial pieces of information like product features and corresponding opinions2,3 expressed on them. In order to identify and extract these linguistic patterns from the reviews, the concept of opinion mining4 is useful. Opinion mining mainly focuses on opinion the text expresses.

Feature level opinion mining5 concentrates on directly extracting the opinion targets themselves, the related opinions and discovers what exactly consumers like or dislike about the opinion target. Often, the feature level opinion mining is specific to the problem under analysis. In order to extract maximum number of features from the product reviews, a comprehensive feature extraction approach is needed which explores the natural language rules for the extraction of various kinds of product features from the reviews.

2. Existing Work

Feature level opinion mining is considered as a major research work for more than a decade. Quite a number of researchers have focused their research on this particular subject. Some of the most outstanding research works are reviewed in this section.

2.1 Explicit Features Extraction

The frequent features6 were extracted using Apriori algorithm7 from the product reviews. The frequent features were extracted by learning the relationship patterns8 among the product feature in the analysis. RedOpal system was developed9 to find the products based on the extracted frequent features. The drawbacks with these approaches are that these approaches produce many non-features and also miss low frequent features. The idea of dependency relations among the product features and opinions in the review sentences was explored10. The unigram product features are only extracted. An improvement in the work on dependency relations was performed wherein the phrases in the reviews are analyzed for bigram product features11. Some of these bigram product features are observed to be the relevant features. An algorithm was proposed12 to extract the product features and opinions in the simultaneous manner. The algorithm was still improved13 to extract the product features using direct and indirect dependency relations among the product features. The major limitation with these approaches is that they produce many non features matching the learned linguistic rules.

2.2 Implicit Features Extraction

In least percentage of online reviews the product feature were expressed in the indirect manner. These features are called as implicit features. The feature indicators which are present in these reviews help to identify the implied product feature. Less amount of research was taken place on identifying implicit features when compared with explicit features. The brief review on extracting the implicit features is discussed below as a survey.

COP-Means clustering algorithm was utilized14 to cluster and link the compatible product feature and opinion words. The implicit features were identified from these clusters by finding out the unlinked feature words in the product feature cluster. A novel co-occurrence association based method was proposed15 to extract implicit features from the customer reviews. This is carried out by calculating the conditional probability of the candidate feature words on the associated notional words. A candidate feature word is considered as implicit feature when the conditional probability of that candidate feature word among others is high. The drawback in implicit feature extraction methods is that they never identified the implicit feature indicator words and also when identified the corresponding implicit features were not extracted.

The feature extraction techniques are particular to the reviews analysis problem. Some of the works gave attention to the extraction of frequent features, some of them concentrated on the relevant features and very few on identifying and extracting implicit features. A comprehensive method for extracting all of these product features from the reviews is required so that the further tasks in opinion mining namely opinion extraction and orientation are carried out in an efficient manner.

3. Extraction of Opinion Targets from Product Reviews using Comprehensive Feature Extraction Model

A domain free approach for comprehensive product feature extraction from online reviews is specified in this section. This approach is based in natural language processing in which the language patterns are identified in each kind of feature extraction. This comprehensive model begins with extracting the frequent features, then finding the relevant features, and next the implicit features and finally extracting the infrequent features. The model is general and is applicable to any domain reviews collection.

Initially, the incoming product reviews are pre-processed. The steps in pre-processing the reviews are explained below.

3.1 Input Reviews Pre-Processing Module

This module is used to pre-process the incoming reviews to a standard format. The steps in pre-processing are namely review tokenization, stop words removal and Part-of-Speech (PoS) tagging. The three datasets namely Rand Mcnally IntelliRoute TND 700 Truck GPS device, Nook Tablet and LCD mounting arm considered for this work are analyzed for preprocessing in this section. These datasets are named as D1, D2 and D3 respectively. The process of review tokenization divides the sentence into individual tokens. Then, the stopwords list is applied on the tokens to remove those words which carry no meaning in the analysis. Massachusetts Institute of Technology (MIT) stop words list containing 570 words are used to remove the stop words from the reviews. This list is appended with 40 additional symbols considered from Stanford CoreNLP package to make the reviews symbol free. Finally, Part of Speech (PoS) tagging is carried out on the list of filtered tokens to unambiguously associate the word category

with each of the token. The Stanford log-linear Part of Speech tagger16 is used for tagging the tokens. The performance of the PoS tagger upon nouns on the three datasets is given in Table 1.

In table 1 it is understood that majority of the words are correctly tagged as nouns. The words wrongly tagged as nouns belong to miscellaneous category. Removal of these nouns from the analysis does not impact the further feature extraction process. The performance of the PoS tagger upon adjectives on the three datasets is given in Table 2.

In Table 2 it is understood that majority of the words are correctly tagged as adjectives. The words wrongly tagged as adjectives belong to miscellaneous category. Removal of these adjectives from the analysis does not impact the further feature extraction process. The PoS tagger suffers with major problems17 namely the unknown words which were not seen in the training phase of the PoS tagger, context level problems in assigning tags and the confusion state of the PoS tagger. The impact of verbs and their variants in opinion orientation of the product features on three datasets is given in Table 3.

In Table 3 it is clear that very less percentage of opinions are identified from verbs and its variants. A majority of the verbs and their variations are not implying any opinion on the product features. Therefore, the verbs are not considered in the further reviews analysis task.

Table 1. PoS tagger performance details on nouns

PoS Details	Statistics on D1	Statistics on D2	Statistics on D3
Number of Nouns (NN and NNS)	379	268	422
Number of adjectives wrongly tagged as noun	5	7	9
Word categories wrongly tagged	MISC	MISC	MISC
% of adjectives wrongly tagged as noun	0.84%	3.7%	2.5%

Table 2. PoS Tagger performance details onadjectives

PoS Details	Statistics on D1	Statistics on D2	Statistics on D3
Number of Adjectives (JJ)	130	72	144
Number of nouns wrongly tagged as adjective	6	2	3
Word categories wrongly tagged	MISC	MISC	MISC
% of nouns wrongly tagged as adjective	3.84%	1.38%	1.38%

Table 3. Impact of verbs in opinion orientation

Dataset/No. of feature, opinion pair identified using verbs	Positive/Negative Impact	% of implied positive/negative opinions
D1/5	2+/3-	0.97% / 1.45%
D2/2	2+/0-	2.4% / 0%
D3/2	2+/0-	0.92% / 0%

Table 4. Final statistics on the PoS tagged datasetwords

PoS Tag	% availability in D1	% availability in D2	% availability in D3
Nouns	99.16%	96.3%	97.5%
Adjectives	96.16%	98.62%	98.62%

The final statistics on the PoS tagged words from the datasets are given in Table 4. The statistics from the above table specify that the product features are possible to identify using the nouns and opinions are possible to identify using adjectives.

Comprehensive Feature ExtractionModule

The step-by-step feature extraction approach is followed to reach the goal of extracting maximum number of product features. Various steps in feature extraction are namely frequent features extraction, relevant features extraction, implicit features extraction and infrequent features extraction. Nouns are extracted as product features as the research18 confirmed that 60-70 % of the features are explicit nouns. After the implementation of every step, the obtained features are added to the list of features so as to assist the count. In all the steps, WordNet is utilized19 to finalize the extracted noun as a dictionary word. The proposed model is illustrated in Figure 1.

Frequent Nouns Extraction

In general, a review sentence is the combination of a noun phrase and an adjective phrase. This sub module calculates the frequency count of each noun from the nouns and noun phrases which were earlier tagged by the PoS tagger. A noun is regarded as frequent if its occurrence in the reviews is within the three percent from the set of nouns that are found. The obtained frequent nouns are stored in a file and are used for further analysis.

Relevant Nouns Identification

The nouns which are written less in number in online reviews are relevant nouns and infrequent nouns respectively. The relevant nouns specify the associated information on the

actual features of the product. A closer analysis of the reviews corpus revealed interesting clues for identifying the relevant nouns. These are specifically, the nouns that are modified by multiple adjectives, the part-whole relation patterns among the product fea- tures, and the adjectives modifying the frequent nouns. The collection of the adjectives that are available adjacent to the nouns and frequent nouns is carried out.

Once these adjectives are collected, the corresponding nouns are extracted as relevant nouns. The PoS patterns that are learned for extracting the relevant nouns are given below:

word_JJ word_NN, word_JJ word_NN word_NNS

Also, the sub-features of the actual features are also extracted as relevant nouns. The obtained relevant nouns are added to the set of frequent nouns which are extracted and stored in the previous step for further analysis.

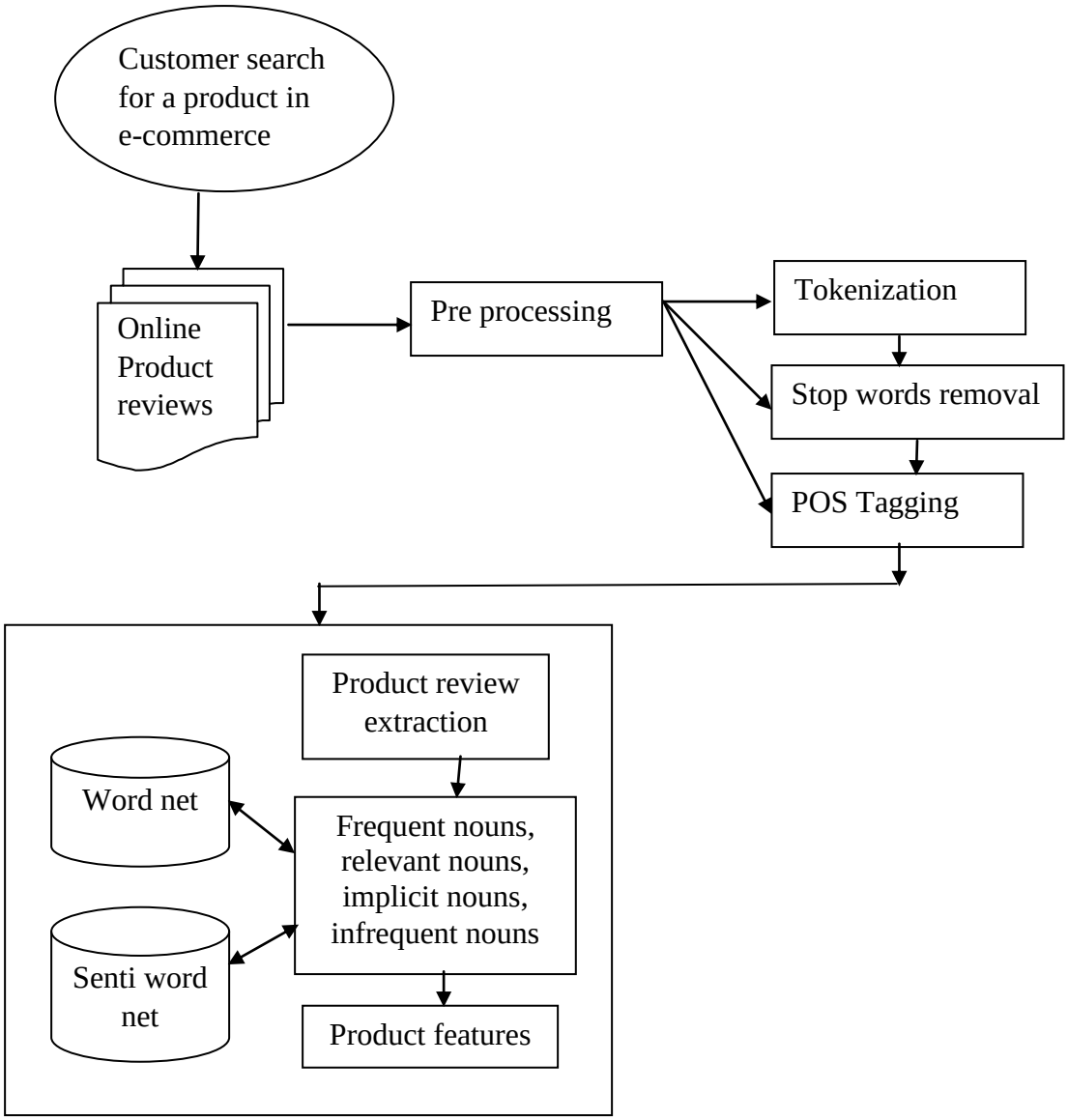

Figure 1. Proposed model.

Implicit Nouns Identification

In some of the reviews, the product features are not writ- ten in an explicit manner. The features in such reviews are called as implicit features. The nouns pertaining to these features are called as implicit nouns. The identification of these nouns is a complex task. In order to carry out this task, the feature indicators which are present in the implicit featured reviews are identified, and with the help of SenticNet knowledgebase20, the nouns specific to the identified feature indicators are extracted. The identification of implicit feature indicators is performed using the Conditional Random Field (CRF)21 sequence labeling model based CRF++ framework. Similar kind of work on identifying the implicit feature indicators was carried out22 in their work. The obtained implicit nouns are also added to the previous list of frequent nouns and relevant nouns and are used for further analysis.

Infrequent Nouns Extraction

As specified earlier, the infrequent nouns are also pres- ent less in number in the online reviews. These nouns are found to be interesting for certain section of customers who want to purchase the product. A noun is regarded as infrequent if its occurrence in the reviews is less than three percent from the set of nouns that are found. The obtained infrequent nouns are stored finally in the previously updated file. The updated file with all the kinds of nouns is considered as the product features.

4. Experimental Results andDiscussion

The electronic product reviews corpus which was obtained from Amazon is used for this experiment. This corpus consists of eleven consumer product reviews.

Table 5. Dataset details

Document attributes	Values
Number of review documents	1900
Minimum sentences per review	1
Maximum sentences per review	25
Minimum number of words per review sentence	25
Maximum number of words per review sentence	32.8

Table 6. Information retrieval measures at each step of feature extraction

Datasets	**Precision (%)**				**Recall (%)**				**F1-score (%)**			
	FN	**RN**	**IN**	**Inf. N**	**FN**	**RN**	**IN**	**Inf. N**	**FN**	**RN**	**IN**	**Inf. N**
D1	70	72.2	75.8	77.9	46.6	53.3	62.6	70.6	55.9	61.3	68.5	72.8
D2	72.3	76.7	79.5	80.8	70	70.6	77.3	84	71.1	73.5	78.3	82.3

D3	65.6	68.4	78.5	83.3	54.2	61.3	72	78.6	59.3	64.6	75.1	80.8

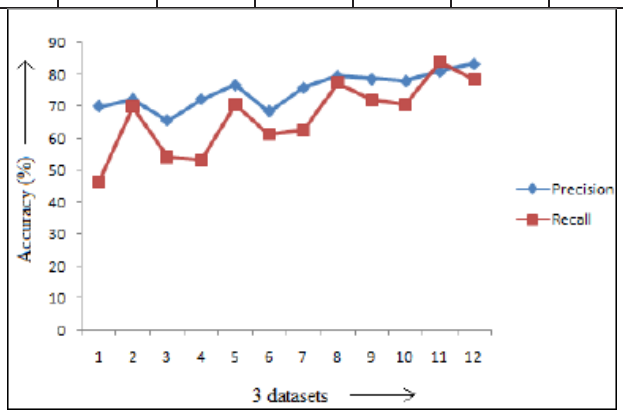

Figure 2. Accuracy of various extracted product features.

Three product reviews were considered for conducting this experiment. Rand Mcnally IntelliRoute TND 700 Truck GPS device, Nook Tablet and LCD mounting arm are the products for which the reviews were considered for analysis. The labels provided to these three datasets are D1, D2 and D3. Table 5 presents the details of the dataset used for this experiment.

The pre-processing of the reviews is already carried out in Section 3 of this work. The evaluation of the extracted features is carried out at each step in the extraction pro- cess. To analyze the performance of the proposed method, the standard Information Retrieval (IR) measures namely Precision, Recall and F1-score are used. The evaluation measures calculated at every step is tabulated in Table 6.

Table 6 shows the performance values of the pro- posed feature extraction model. The evaluation process starts by performing the task of extracting greater than or equal to 3% of most frequent nouns from the reviews of each product. This task extracted product features with acceptable levels of precision on the three datasets. There is a decrease in the recall which is observed on the three datasets. Further, the relevant nouns are extracted from the reviews by the nouns that are modified by multiple adjectives, the part-whole relation patterns among the product features, and using the adjectives identified on the frequent nouns. It is observed that there is an increase in the precision after implementing this step with an average increase in the recall on the three datasets. The next step of discovering implicit nouns from the reviews were also observed a good increase in the precision as the average number of identified implicit nouns from the collected reviews dataset is 57%. There observed a significant increase in the recall after extracting implicit nouns.

Finally, the task of extracting infrequent features is implemented by applying the reverse condition of frequent features. This task considerably increased the precision of the three product datasets. It is observed that the recall on the three datasets has been increased in a considerable manner. This last step of extracting infrequent features ensures maximum

retrieval of the product features. The average percentage of irrelevant product features across the three datasets is 17%. The results are shown in Figure 2.

By achieving an average precision of above 75% across all the steps of feature extraction on all the three datasets, it is concluded that the comprehensive feature extraction approach performs better than the particular way for extracting the product features in the semantic environment.

5. Conclusions and Future Work

The extraction of all kinds of explicit product features and implicit product features using the comprehensive feature extraction model was carried out successfully. The objective is to extract the maximum and exact product features from a large number of online product reviews. The experimental results indicate that the proposed model is effective.

In future, the opinions of the extracted product features are to be identified. These identified features and opinions are analyzed for feature specific intentions. These intentions are useful in recommending the similar products in a better way than the traditional recommendations when a search for particular product takes place. This advanced data model helps the businesses to decrease their customer churn.

6. References

1. What Is Web 2.0. Design patterns and business models for the next generation of software. 2016. Crossref

2. Santosh DT, Vardhan BV. Automatic machine recognition of features and sentiments from online reviews. IAENG Transactions on Engineering Sciences. Special Issue for the International Association of Engineers Conferences; 2015. p. 264–80.

3. Santosh DT, Babu KS, Prasad SD, Vivekananda A. Opinion mining of online product reviews from traditional LDA Topic Clusters using Feature Ontology Tree and Sentiwordnet. IJEME. 2016; 6:1–11.

4. Liu B. Opinion mining. Encyclopedia of Database Systems; Springer US. 2009. p. 1986–90.

5. Liu B, Zhang L. A survey of opinion mining and sentiment analysis. Mining Text Data; Springer US. 2012. p. 415–63.

6. Hu M, Liu B. Mining opinion features in customer reviews. AAAI. 2004 Jul; 4(4):755–60.

7. Agrawal R, Srikant R. Fast algorithms for mining asso- ciation rules. Proc 20th Int Conf. Very Large Data Bases, VLDB. 1994 Sep; 1215:487–99.

8. Popescu AM, Etzioni O. Extracting product features and opinions from reviews. Natural Language Processing and Text Mining; Springer London. 2007. p. 9–28. Crossref

9. Scaffidi C, Bierhoff K, Chang E, Felker M, Ng H, Jin C. Red Opal, product-feature scoring from reviews. Proceedings of the 8th ACM Conference on Electronic Commerce; ACM. 2007 Jun. p. 182–91.

10. Zhuang L, Jing F, Zhu XY. Movie review mining and sum- marization. Proceedings of the 15th ACM International Conference on Information and Knowledge Management; ACM. 2006 Nov. p. 43–50. Crossref

11. Wu Y, Zhang Q, Huang X, Wu L. Phrase dependency parsing for opinion mining. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2009; 3:1533–41. Crossref

12. Wang B, Wang H. Bootstrapping both product features and opinion words from Chinese customer reviews with cross- inducing. IJCNLP. 2008 Jan; 8:289–95.

13. Qiu G, Liu B, Bu J, Chen C. Opinion word expansion and tar- get extraction through double propagation. Computational linguistics. 2011 Mar; 37(1):9–27. Crossref

14. Su Q, Xu X, Guo H, Guo Z, Wu X, Zhang X, Swen B, Su Z. Hidden sentiment association in Chinese web opinion mining. Proceedings of the 17th International Conference on World Wide Web; ACM. 2008 Apr.
p. 959–68. Crossref

15. Zhang Y, Zhu W. Extracting implicit features in online cus- tomer reviews for opinion mining. Proceedings of the 22nd International Conference on World Wide Web; ACM. 2013 May. p. 103–4. Crossref

16. Toutanova K, Klein D, Manning CD, Singer Y. Feature- rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Association for Computational Linguistics. 2003; 1:173–80. Crossref

17. Manning C. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? Computational Linguistics and Intelligent Text Processing; 2011. p. 171–89.

18. Liu B. Web data mining, exploring hyperlinks, contents and usage data. Springer Science and Business Media; 2007. p. 532.

19. Kilgarriff A, Fellbaum C. WordNet. An Electronic Lexical Database. Cambridge, MA. 1998; 422:1–7.

20. Biagioni R. The SenticNet sentiment lexicon: Exploring semantic richness in multi-word concepts; Springer. 2016 May. p. 64.

21. Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the Eighteenth International Conference on Machine Learning, ICML. 2001 Jun; 1:282–9.

22. Poria S, Cambria E, Ku LW, Gui C, Gelbukh A. A rule- based approach to aspect extraction from product reviews. Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP); 2014 Aug.
p. 28–37. Crossref