Feature and Sentiment based Linked Instance RDF Data towards Ontology based Review Categorization

D. Teja Santosh, B. Vishnu Vardhan, *Member, IAENG*

Abstract-Online reviews have a potential impact on the green customer who wants to purchase or consume the product through e-commerce. Online reviews contain features which are useful for the analysis in opinion mining. Most of the today's systems work on the summarization of the features taking the average features and their sentiments leading to structured review information. Often the context of surrounding feature is undermined which helps while classifying the sentiment of the review. In web 3.0 machine interpretable Resource Description Framework (RDF) were introduced which helps in structuring these unstructured reviews in the form of features and sentiments obtained from traditional preprocessing and extraction techniques. The context data also supports for future ontology based analysis taking support of Wordnet lexical database for word sense disambiguation and Sentiwordnet scores used for sentiment word extraction. Many popular RDF vocabularies are helpful in the creation of such machine processable data. In the future work, such instance RDF data will be used in the OWL Ontology to reason the data to clearly identify the features and sentiments against the applied data set. These results are sent back to the interface as corresponding {feature, sentiment} pair so that reviews are filtered clearly and helps in satisfying the feature set of the customer.

Keywords-Opinion mining, Feature, Sentiment, Resource Description Framework, Ontology

I. INTRODUCTION

User involvement in writing online reviews for the experience on a specific product is increasing now a days and it is a driving factor in purchase decision making. In web 2.0, the social web is introduced and with this provision, its database is increased from time to time leading to the plethora of reviews. These reviews which are regularly fed into the site are not useful for certain cross section of people. This has led to the concept of Opinion Mining [2].

D. Teja Santosh is the Research Scholar in the Department of Computer Science and Engineering at Jawaharlal Nehru Technological University, Kakinada, India (e-mail: tejasantoshd@gmail.com).

B. Vishnu Vardhan is the Professor in the Department of Computer Science and Engineering at Jawaharlal Nehru Technological University, University College of Engineering, Jagityal, Karimnagar, India (e-mail: mailvishnu@yahoo.com).

The requirement for categorizing reviews on the basis of extracted features with corresponding sentiments is potentially increased in recent past. The obtained features and sentiments are used in converting the unstructured reviews into a form which is suitable for data analysis task. Semantic Web's Resource Description Framework (RDF) [3] is used to structure the review data useful for opinion mining. Major RDF vocabulary metadata are used in the creation of such RDF. Generally RDF allows data to be processed outside the environment where it was created. SPARQL queries can be targeted on the RDF data to validate the features and sentiments on the reviews. This forms the basis for creating a standard OWL Ontology [4] which is used as the structured data model (knowledge model) with rich semantics towards identifying feature based review categories. The obtained categorizes filter the reviews making the purchase decision faster and accurate.

II. RELATED WORK

Sentiment Analysis of online reviews has received major research work in identifying features and extracting the sentiment/opinion words. Minqing Hu and Bing Liu's [5] work on Mining and Summarizing Customer Reviews in which various product features are identified using Apriori Association Algorithm technique. Then they were used Bipolar Adjective Structure in identifying the opinion words and its role in opinion orientation. Shitanshu Verma and Pushpak Bhattacharyya [6] worked on SentiWordNet lexical resource to extract sentiment attached with the features in the review.

Machine Learning techniques are limited in classifying the online reviews based on the binary polarity classes i.e., positive or negative. Research in this aspect has also gained importance for improving the machine's performance for future unseen reviews. Christopher C. Yang et al. [7] used naïve Bayesian classifier as Machine Learning algorithm and used only those features obtained from Information Gain for sentiment classification.

Ontology based Opinion Mining has also been researched extensively in literature. Larissa A. de Freitas and

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Renata Vieira [10] were extracted hotel and movie features from respective Ontologies and summarized the features. Jantima Polpinij and Aditya K. Ghose [8] classified the sentiments using lexical variable ontology for identifying features and used SVM classification for sentiment classification and achieved 96% classification accuracy. Rafael Valencia-Garcia et al. [9] worked on Movie reviews using Movie Ontology and extracted features and used Geometric polarity pyramids in determining opinion words.

In the process of understanding the Ontology based surveys, certain shortcomings are identified: First, context of the review were not considered as the reviews are written based on the experience of the consumers feature set. This experience varies accordingly with consumer to consumer. Also, the context information helps in clearly disambiguate the sense of the review thereby leading to more clear sentiment classification of the reviews. Second, the opinion mining was not progressed further while summarizing the features of the reviews. This can be extended towards web 3.0 or semantic web based reviews analysis using RDF. Ontology can be constructed for effective reasoning of the reviews. Finally, Machine Learning Algorithms are used on the generated sentiment data in web 2.0 works. These algorithms can also be applied to Ontology which can be considered as a domain based data model rich in semantics with data in obvious structured form (used as a standard web 2.0 data model) for both Sentiment Classification and feature categorization.

III. FEATURE AND SENTIMENT BASED LINKED INSTANCE RDF DATA

The principle objective of the approach proposed in this paper is to convert the unstructured reviews to a structured data using RDF and the obtained features and sentiments are to be further refined towards Ontology for Opinion mining. In order to achieve this goal, a framework presented in Fig.1 which is composed of three main modules such as Natural Language Processing (NLP) as an input preprocessing module, Feature Extraction and Sentiment Orientation and Sentiment word Extraction module and the RDF Instance DB Creation module. A detailed description of these components is provided below.

A. NLP Module

This component is used to normalize or preprocess the incoming review mainly focusing on the logical restructuring of the review to a standard format. Text normalization involves operations like identifying processing tokens by tokenization with special symbols requiring special processing. Stop Words Algorithm is applied to eliminate those tokens that have little value to the system, identifying specific word characteristics using Morphological Analysis and part-of-speech tagging for assisting disambiguation of a particular word.

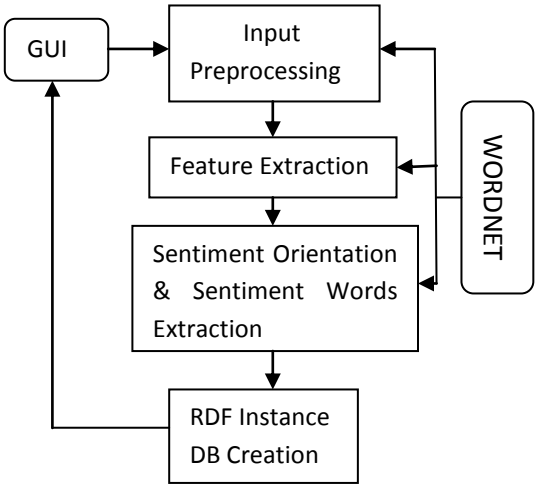

Fig. 1 Proposed System Architecture

B. Feature and Sentiment Orientation module

1.1 Feature Extraction

In Opinion mining, feature extraction is one of the most complex tasks, since it requires the use of Natural Language Processing techniques in order to automatically identify the features in the opinions under analysis. The Feature Extraction process receives an input as a text containing an opinion, and returns the extracted features. The extraction process includes four main steps namely frequent nouns identification, relevant nouns identification, context dependent Features identification, and irrelevant feature pruning.

1.1.1 Frequent Nouns Identification

Many of the product reviews contains the nouns. These are found to be the features of a particular product. A noun is considered to be frequent if it's occurrence in the reviews is at least greater than or equal to three percent [13] from the set of nouns found.

1.1.2 Relevant Nouns Identification

The frequent features obtained from the above step are still found to have fewer relevancies. Adjectives adjacent to the frequent nouns are considered to be the main features where these adjectives are specified on the noun part of the sentence. Now, the obtained new features are added with the existing set of frequent features to make a relevant set of features.

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

1.1.3Context Dependent Features Identification

In some of the reviews the adjectives specified are highly domain dependent. These adjectives and to some extent the adverbs refer to the feature present in the review implicitly. Identifying such features is a complex task. This is carried out by mapping individually with the available list of such adjective and adverb combination features. When these combinations are identified, then the corresponding features are updated to the feature set.

1.1.4Irrelevant feature pruning

The feature set contains irrelevant features such as the features which are uninteresting. These features are to be removed or pruned from the feature set. This is processed by identifying the feature phrases which are not strongly connected together among the review sentences.

1.1.5Synonym Grouping

The similar product features can be expressed in different words. In order to have genuine features in the set, grouping is made for the feature synonyms.

1.2 Sentiment Orientation and Sentiment Words Extraction

Sentiment Orientation or opinion word Orientation reflects to either the positive or negative side of the review. Sentiment annotation to the words is carried out by using a popular dictionary based approach named as WordNet [11] information. Following are the steps performed to extract the sentiment from a specific review.

1. Select one of the five random five-word subsets of some Harvard Inquirer oppositions.
2. Provide weight values to these words using SentiWordNet [12]. Also, fix the seed terms as "good" and "bad" for the two seed sets.
3. Process all the adjectives from the reviews and make two sets as Positive Adjectives set (PA) and Negative Adjectives set (NA) and map the synonyms from WordNet and review adjectives. This completes the extraction of sentiment words.
4. Calculate the Sentiment Orientation (SO) of a term 't' taken from PA and NA by its relative distance from the two seed terms good and bad as;

$$SO(t) = \frac{dist(t, bad) - dist(t, good)}{dist(good, bad)}$$

where 'dist' is the measure between two terms t1 and t2

5. The given term is deemed to be positive if Orientation measure is greater than zero and negative otherwise.

For the analysis, one of the positive and negative seed sets is selected as following (Turney & Littman, 2003):

S_p = {good, nice, excellent, positive, fortunate, correct, superior}
S_n = {bad, nasty, poor, negative, unfortunate, wrong, inferior}

When required map the extracted adjectives with the corresponding seed set adjectives. The score for the adjective from the given review is 0.125 for "new" (negative word) from SentiWordNet [12] and a score of 1 for "good" and 0 for "bad". The Orientation measure for the term is 1. But, this doesn't mean that the Sentiment expressed by the review is positive because the sentiment value is calculated based on its proximity to the feature.

C. RDF Instance DB Creation module

Resource Description Framework (RDF) is a framework for the representation of resources. RDF is a data model originally used for metadata for the web resources. Over the web, a resource can be anything that is located via a URI (Uniform Resource Identifier). The basic building block for RDF creation is the statement which can be represented in the form of triples (subject, predicate and object in a sentence form a triple). RDF uses a graph data model where different entities are vertexes in the graph and relationships between them are represented as edges. Information about an entity is represented by directed edges emanating from the vertex for that entity (labeled by the name of the attribute), where the edge connects the vertex to other entities, or to special literal vertexes that contain the value of a particular attribute for that entity. Linked data is the web of data linked universally in order to better understand the entities present over the web. Millions of triples till date are connected making the Web as "Web of Data" rather than "Web of Documents". With RDF, any relational data can be represented as triples. The mapping from Relational Database to RDF can be done as following:

- Row Key in RDB corresponds to Subject in RDF.
- Column in RDB corresponds to Predicate in RDF.
- Value in RDB corresponds to Object in RDF.

MARL Ontology vocabulary is chosen as the appropriate vocabulary as it can enable to publish data about opinions in the form of linked data efficiently. The extracted features and the sentiments are populated with MARL vocabulary. With

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the mappings specified in (1), conversion of unstructured reviews to structured forms like RDF and RDB table is easy. The triple in Table 1 generated after validating the RDF is given below.

Table 1 RDF Triple

Statement	Value
Subject	http://goo.gl/8OL3sO
Predicate	rdf:type
Object	marl:Opinion

rdf:type is an instance of rdf:Property that is used to state that a resource is an instance of a class. The predicate "type" links the *about* data: product URI (subject) with the object "Opinion".

The context present in the review is the key to the overall sentiment of the review. At feature level opinion mining, context is understood as the clue to the feature present in the review to disambiguate its sense.

"The ***chair*** emphasized the need for the education"

In the above sentence, '*emphasized*' is the clue to say that chair is a person. This is crucial information which can clearly classify the reviews based on the selection of the feature and its corresponding synonym group. Now, the updated table for the considered review is;

Table 2 Opinion RDB

describesFeature	hasContext	hasPolarity	Rating
Design	Phone	Positive	5

IV.RESULTS

Following the steps in extracting features and sentiments as explained in section 3, those are tabulated step wise. Also, a comparative histogram is generated in Fig. 3 which shows both the positive and negative sentiments on the features obtained.

Table 3 Feature data obtained from methodology of Nokia 6600 reviews

Frequent Features	Relevant Features	Context dependent Features	Features are Irrelevant Features Pruning	Synonyms Groups Names set
{design, image, picture, zoom, flash, size, battery, powerup, quality, camera, LCDScreen, DigicIIchip, print}	{design, image, picture, zoom, flash, size, battery, powerup, quality, camera, LCDScreen, DigicIIchip, print, *digitalcard, autofocus, processor, camerashake*}	{design, image, picture, zoom, flash, size, battery, powerup, quality, camera, LCDScreen, DigicIIchip, print, ***digitalcard, autofocus, processor, camerashake,*** *dslsr, SD card*}	{design, image, picture, zoom, flash, size, battery, powerup, quality, camera, LCDScreen, DigicIIchip, print, ***digitalcard, autofocus, processor, dslsr, SD card***}	{Image, Memorycard, Processor}

A good number of triples in RDF notation make the retrieved review features and sentiments important as these can be used further in machine learning cross validation analysis. As compared to manual extraction of features from [5], the number of triples is 402. By combining both the triple counts a total of 4402 triples. This count can help in categorizing the reviews in a better way.

Table 4 Sentiment data obtained from methodology of Nokia 6600 reviews

Sentiment Orientation	Sentiment Words
Positive	{large, bright, big, best, brilliant, awesome, great, new, easy, cool, worth, decent, polite, terrific, nice}
Negative	{steep, slow, excessive, bulky, terrible, worst, difficult, awful, bad}

The RDF triples are stored in Sesame RDF Repository presented in Fig. 2 and can be queried over the Feature and Sentiment linked data using SPARQL.

Fig. 2 Sesame RDF Repository showing feature and sentiment linked data contexts

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V.CONCLUSION AND FUTURE WORK

The feature and sentiment based linked instance RDF data has been done successfully. Population into RDF database helps in classifying the reviews with the help of Product Review Opinion Ontology (PROO) OWL Ontology constructs and machine learning algorithm. The so obtained categories help the customer in taking wise decisions for purchasing the product with less time.

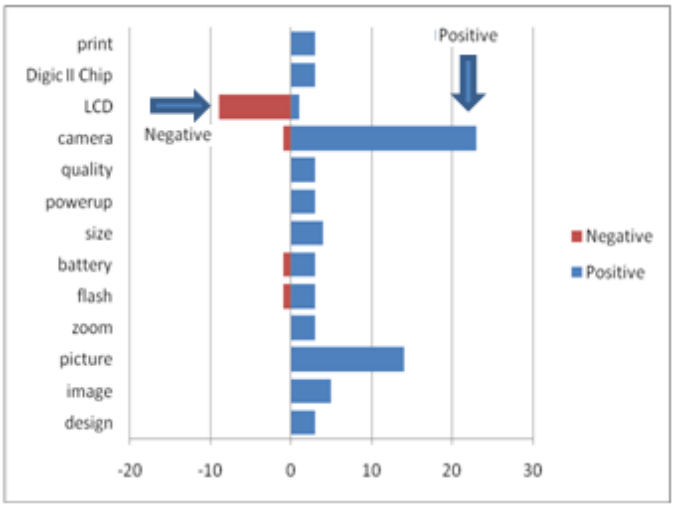

Fig. 3 Comparative Histogram showing the positive and negative sentiments of the features

REFERENCES

[1] Alekh Agarwal and Pushpak Bhattacharyya, *Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified,* Proceedings of ICON, 2005.

[2] Bo Pang and Lillian Lee, *Opinion mining and sentiment analysis,* Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135.

[3] Selver Softic and Michael Hausenblas, *Towards Opinion Mining Through Tracing Discussions on the Web,* 2008.

[4] Paul Buitelaar et al., *Linguistic Linked Data for Sentiment Analysis,* August 2013.

[5] Minqing Hu, *Bing Liu, Mining and summarizing customer reviews*, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22-25, 2004, Seattle, WA, USA.

[6] Verma, S., & Bhattacharyya, P., *Incorporating semantic knowledge for sentiment analysis*, Proceedings of ICON, 2009.

[7] Christopher C. Yang , Y. C. Wong , Chih-Ping Wei, *Classifying web review opinions for consumer product analysis*, Proceedings of the 11th International Conference on Electronic Commerce, August 12-15, 2009, Taipei, Taiwan.

[8] Polpinij, J., & Ghose, A. K., *An ontology-based sentiment classification methodology for online consumer reviews*, Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 518-524). IEEE Computer Society, December 2008.

[9] Peñalver-Martínez, Isidro, Rafael Valencia-García, and Francisco García-Sánchez, *Ontology-guided approach to feature-based opinion mining*, In *Natural Language Processing and Information Systems*, pp. 193-200. Springer Berlin Heidelberg, 2011.

[10] Freitas, Larissa A., and Renata Vieira, *Ontology based feature level opinion mining for portuguese reviews*, In *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 367-370. International World Wide Web Conferences Steering Committee, 2013.

[11] Christiane Fellbaum (1998), *WordNet: An Electronic Lexical Database. Bradford Books*.

[12] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani, *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, In *LREC*, vol. 10, pp. 2200-2204. 2010.

[13] Henrique Siqueira and Flavia Barros, *A Feature Extraction Process for Sentiment Analysis of Opinions on Services,* Proceedings of International Workshop on Web and Text Intelligence. 2010.

ISBN: 978-988-19253-4-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)