International Journal of Technology (2015) 2: 198-206
ISSN 2086-9614

OBTAINING FEATURE- AND SENTIMENT-BASED LINKED INSTANCE RDF DATA FROM UNSTRUCTURED REVIEWS USING ONTOLOGY-BASED MACHINE LEARNING

D. Teja Santosh[1*], B. Vishnu Vardhan[2]

[1] *Department of Computer Science & Engineering, Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India*
[2] *Jawaharlal Nehru Technological University College of Engineering, Kodimyal mandal, Karimnagar Dist. Telangana, India*

(Received: October 2014 / Revised: March 2015 / Accepted: April 2015)

ABSTRACT

Online reviews have a profound impact on the customer or "newbie" who wishes to purchase or consume a product via Web 2.0 e-commerce. Online reviews contain features that form half of the analysis in opinion mining. Most of today's systems work on the basis of summarization, looking at the average obtained features and their sentiments, leading to structured review information being generated. Often, the context surrounding a feature, which helps the sentiment of the review to be classified clearly, is overlooked. The Web 3.0-based machine interpretable Resource Description Framework (RDF) can be used to structure these unstructured reviews into features and sentiments, which are obtained via traditional preprocessing and extraction techniques. Here, data about the context is also provided for future ontology-based analysis, with support from the WordNet lexical database for word sense disambiguation and SentiWordNet scores for sentiment word extraction. Many popular RDF vocabularies are helpful for obtaining such machine-processable data. This work forms the basis for creating/upgrading the (available) OWL Ontology that can be used as a structured data model with rich semantics for supervised machine learning. With this method, the classified sentiment categories are validated in relation to precise sentiments and are sent back to the interface in corresponding "feature/sentiment" pairs so that reviews are filtered clearly, which helps to satisfy the feature set of the customer.

Keywords: Feature; Ontology; Opinion mining; Resource Description Framework; Sentiment

1. INTRODUCTION

User involvement in writing online reviews about their experience of a product is a huge driving factor in purchase decision making. The periodic expansion of this "social web" provision in Web 2.0 has led to a plethora of reviews being available. The reviews, which are regularly fed on to the site, cannot be read by customers in full. This has led to the concept of *opinion mining* (Pang & Lee, 2008). The requirement for the categorization of reviews on the basis of extracted features with corresponding sentiments has thus increased. The obtained features and sentiments are used to convert unstructured reviews into a form suitable for data analysis. Semantic Web's Resource Description Framework (RDF) (Softic & Hausenblas, 2008) can be used to structure review data and is therefore useful for opinion mining. Major RDF vocabulary metadata is used in the creation of such RDF.

* Corresponding author's email: tejasantoshsharma@gmail.com, Tel. +9246214659
Permalink/DOI: http://dx.doi.org/10.14716/ijtech.v6i2.555

Further, RDF allows data to be processed outside the environment in which it was created. SPARQL queries can target the RDF data in order to validate the features and sentiments within reviews.This forms the basis for creating a standard OWL Ontology (Buitelaar et al., 2004), which can be used as structured data model (knowledge model) with rich semantics for machine learning, through which feature-based text categories can be generated. These categorizes filter the reviews, leading to greater speed and accuracy in customers' purchase decision-making processes.

The remainder of the paper is organized as follows: related work is described in Section 2 and the proposed method is explained in Section 3; the way in which the obtained RDF data is utilized as semantic data for machine learning (Agarwal & Bhattacharyya, 2005) using OWL Ontology is then briefly explained in Section 4, alongside a discussion of the results, before conclusions and suggestions for future work are presented in Section 5.

2. RELATED WORK

The sentiment analysis of online reviews has received major academic attention in terms of the identification of features and the extraction of sentiment/opinion words. For example, Hu and Liu (2004) work on mining and summarizing customer reviews examines the frequent features related to various products using the Apriori Association Algorithm technique, as well as employing Bipolar Adjective Structure to identify opinion words and their role in opinion orientation. Verma and Bhattacharyya (2009) use the SentiWordNet lexical resource to extract the sentiments connected with features in reviews.

Machine learning techniques are limited in their ability to classify online reviews into binary polarity classes – i.e., positive or negative. Research in this aspect has also gained importance in the pursuit of improving machines' performance in relation to future unseen reviews. Yang et al. (2009) use the Naive Bayes classifier as a machine learning algorithm, utilizing only those features obtained from Information Gain for sentiment classification.

Ontology-based opinion mining has also been researched extensively in the literature. de Freitas and Vieira (2013) extract hotel and movie features from respective ontologies and summarize their features. Polpinij and Ghose (2008) classify sentiments using lexical variable ontology to identify features and SVM classification for sentiment classification, achieving 96% classification accuracy. Peñalver-Martínez et al. (2011) analyze movie reviews using Movie Ontology to extract features and Geometric Polarity Pyramids to determine opinion words.

In this body of literature, the following shortcomings are identified: first, the context of the reviews is not considered, even though the reviews are written based on the experience of the consumers with their feature set. This varies among the consumers. Context information aids in the disambiguation of the sense of a review, thereby leading to more accurate sentiment classification. Second, in these studies, opinion mining was often limited to the feature-based summarization of the reviews. It could be extended to encompass Web 3.0-based or semantic web-based review analysis using an RDF instance DB and an existing ontology or newly identified concepts and properties in order for effective reasoning to be achieved. Finally, existing work tends to apply machine larning algorithms to the generated sentiment data in Web 2.0 works. These algorithms can also be applied to ontologies, which are domain-based data models that are rich in semantics and in which the data is in an obvious structured form (using a standard Web 2.0 data model), thus allowing sentiment classification and feature categorization to take place. These feature categories can then be compared with reasoned conclusions drawn from ontologies.

3. FEATURE- AND SENTIMENT-BASED LINKED RDF DATA

The principal objective of the approach proposed in the manuscript is to convert unstructured review text into structured data using RDF to obtain features and sentiments, which are then further refined by using ontology for opinion mining. In order to achieve this goal, a framework (see Figure 1.) is provided, composed of 3 main modules: Natural Language Processing (NLP) as an input preprocessing module, a feature extraction and sentiment orientation module, and the RDF Instance DB Creation module. A detailed description of these components is provided below.

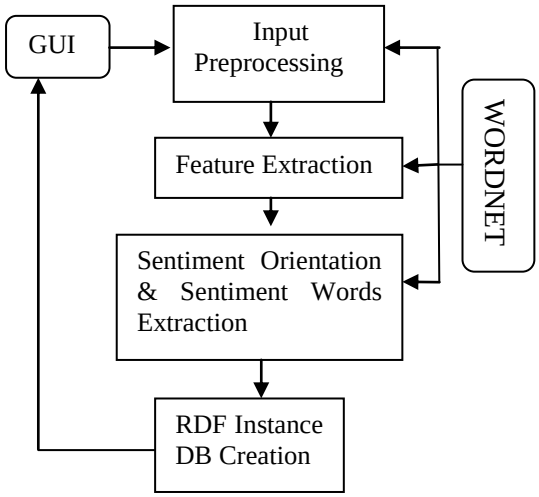

Figure 1 Proposed system architecture

3.1. NLP Module

This component is used to normalize or preprocess an incoming review – i.e., for the logical restructuring of the review into a standard format. Text normalization involves operations like tokenization, stop word algorithms, and Part of Speech (PoS) tagging. This aids accurate semantic analysis at a later stage. First, all punctuation marks are removed from the review. For example, the only punctuation mark available in the review "The camera was sitting on my home desk waiting for the new digital card and the weekend for the first round of serious shooting." is the period (.). This is removed from the review. Then, the review is tokenized, converting sentences into sequences of words. This prevents sentences from being treated as sequences of characters. The tokens are listed in a set, such as "camera, was, sitting, on, my, home, desk, waiting, for, the, new, digitalcard, and, the, weekend, for, the, first, round, of, serious, shooting" in this example.

Second, a stop-word list is applied to filter the obtained tokens, eliminating those which feature infrequently in the search process. The Stanford NLP Application gives the following set of stop words: "a, an, and, are, as, at, be, by, for, from, has, in, is, it, its, of, on, that, the, to, was, were, will, with." The resulting set of words in our example is: "camera, sitting, my, home, desk, waiting, new, digitalcard, weekend, first, round, serious, shooting."

Next, morphological analysis is undertaken on the tokens to identify the morphemes present. Morphological analysis assists in the identification of root words in a language, which will make clearly understand the words derived from it for context oriented processing. Also, it is useful for Part of Speech (PoS) tagging, helping to disambiguate word senses. In the given review, there are four words formed from five root words: "sit," "wait," "week," "end," and "shoot." The words formed from these root words are called *bound morphemes*. The set "sitting, waiting, shooting" is composed of gerund verbs, which are identified and tagged as

noun forms because they are formed by adding "–ing" to the root word. There are also six *free morphemes* present: "home," "new," "digitalcard," "first," "round," and "serious." Morphology trees for the bound morphemes are illustrated below.

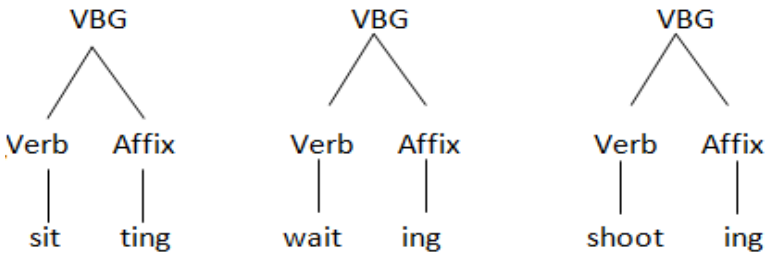

Figure 2 Morphological trees

Subsequently, lexical analysis is performed via PoS tagging to disambiguate particular tokens or words senses. This tagging helps to extract features from reviews (nouns and, to some extent, from attached adjectives and adverbs) using the properties of the words present. A PoS-tagging set of tools from the Stanford Log-linear Part-Of-Speech Tagger framework is used to perform the major task of tokenization.

3.2. Feature and Sentiment Orientation Module

3.2.1. Feature extraction

In opinion mining, feature extraction is one of the most complex tasks, since it requires the use of Natural Language Processing techniques to identify the features present in the opinions under analysis automatically. In the feature extraction process, text containing an opinion is inputted the extracted features are returned.

3.2.1.1. Frequent nouns identification

The majority of product reviews contain nouns. These nouns tend to refer to the features of a particular product. A noun is considered to occur frequently if it appears in a review at least three times.

3.2.1.2. Relevant nouns identification

The frequent features obtained via the previous step are not necessarily the most relevant. Adjectives adjacent to the frequent nouns are considered to be important features as the adjectives are specified on the noun part of the sentence. The new features obtained here are added to the existing set of frequent features to make a relevant set of features.

3.2.1.3. Context-dependent features identification

In some reviews, the adjectives specified are highly domain dependent. These adjectives and, to some extent, the adverbs refer implicitly to the features present in a review. Identifying such features is a complex task. This is achieved by mapping individually with the already available list of such adjectives and adverbs combination of features. If such features are identified, then they are added to the feature set.

3.2.1.4. Irrelevant feature pruning

The feature set might contain irrelevant features – i.e., features that are uninteresting to the researcher. These features can be removed or "pruned" from the feature set. This requires a process of identifying the feature phrases that are not connected strongly (compacted together) within the review sentences.

3.2.1.5. Synonym grouping

The same product features can be expressed using different words. In order to ensure that only genuine features appear in the set, synonymous features are grouped together.

3.2.2. Sentiment orientation and sentiment word extraction

Sentiment orientation (or *opinion word orientation*) determine the positive and/or negative sides of a review. Sentiment annotation can be performed by using a popular dictionary-based approach that uses the WordNet database (Fellbaum, 1998). The following steps are performed to extract sentiments from a review:

1. One out of five random five-word subsets of Harvard Inquirer oppositions is selected.
2. Weight values are provided for these words using SentiWordNet (Baccianella et al., 2010). The seed terms are set as "good" and "bad" for the two seed sets.
3. All the adjectives in the reviews are identified and two sets are created: 1) Positive Adjectives (PA) and 2) Negative Adjectives (NA). The synonyms obtained from WordNet and the reviews are then mapped. This completes the extraction of sentiment words.
4. The sentiment orientation (SO) of a term "t" is taken from PA and NA divided by its relative distance from the two seed terms, "good" and "bad," is calculated:

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

   where "dist" is the measurement between two terms, "t1" and "t2."

5. The given term is deemed to be positive if the orientation measurement is greater than zero, and negative otherwise.

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

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

If needed, the extracted adjectives are mapped on to the corresponding seed set adjectives. The score for an adjective from a given review is 0.125 for "new" (negative word) from SentiWordNet (Baccianella et al., 2010) and a score of 1 for "good" and 0 for "bad." The orientation measurement for the term is 1. This does not mean, however, that the sentiment expressed by the review is positive because the sentiment value is calculated based on its proximity to the feature. The phrase from the review ***"waiting for the new digital card and the weekend"*** tells us that no experience had been gained of using the camera in our example, leading to the review being neutral.

3.3. RDF Instance DB Creation Module

The Resource Description Framework (RDF) is a language for the representation of resources. It is a data model that was originally used for metadata for web resources. On the Internet, a "resource" refers to anything that can be located via a Uniform Resource Identifier (URI). The basic building block for RDF creation is a statement that can be represented in the form of a "triple" (the subject, predicate, and object in a sentence create a triple). RDF uses a graph data model in which different entities are vertexes on a graph and the relationships between them are represented as edges. Information about an entity is represented by directed edges that emanate from the vertex for that entity (labeled with the name of the relevant attribute), and the edges connect the vertex to other entities or to special literal vertexes that contain the values of

particular attributes for entities. Linked data evolves over time is the web of data linked universally in order to better understand the entities present over the web. In this way, data becomes structured, allowing very specific queries to be applied, leading to interesting answers being obtained, which current systems cannot do. Millions of triples can be connected in this manner, making the Internet into a web of data, rather than a web of documents.

With RDF, any relational data can be represented as triples or as RDF Statements. The mapping from a Relational Database to RDF can be achieved as follows:

- A row in an RDB corresponds with a subject in RDF.
- A column in an RDB corresponds with a predicate in RDF.
- A value in an RDB corresponds with an object in RDF.
  (1)

MARL Ontology vocabulary is chosen as the appropriate vocabulary as it can enable the efficient obtainment of data about opinions in the form of linked data. The extracted features and sentiments are populated using MARL vocabulary. Utilizing the mappings specified in (1), the conversion of unstructured reviews to structured forms like RDF and the RDB table is easy. Below is the RDF code in MARL vocabulary for one of the extracted features and sentiment data examples:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:marl="http://purl.org/marl/ns#"
        xmlns:v="http://rdf.data-vocabulary.org/#">
<marl:Opinion rdf:about="http://img315.imageshack.us/img315/3918/nokia6600288ju.jpg">
   <marl:extractedFrom
rdf:resource="http://www.amazon.co.uk/review/R1DIIVQ8WRR7OB/"/>
   <marl:describesFeature>design</marl:describesFeature>
   <marl:hasPolarity>Positive</marl:hasPolarity>
   <v:rating>5</v:rating>
</marl:Opinion>
</rdf:RDF>
```

“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.” In MARL vocabulary, “Opinion” is a class. The mapping RDB table for the Nokia 6600 review’s content, with feature and sentiment (polarity) values, is given below, as are the RDF statements generated after validating the RDF:

Number	Subject	Predicate	Object
1	http://img315.imageshack.us/img315/3918/nokia6600288ju.jpg	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://purl.org/marl/ns#Opinion
2	http://img315.imageshack.us/img315/3918/nokia6600288ju.jpg	http://purl.org/marl/ns#extractedFrom	http://www.amazon.co.uk/review/R1DIIVQ8WRR7OB/
3	http://img315.imageshack.us/img315/3918/nokia6600288ju.jpg	http://purl.org/marl/ns#describesFeature	"design"
4	http://img315.imageshack.us/img315/3918/nokia6600288ju.jpg	http://purl.org/marl/ns#hasPolarity	http://purl.org/marl/ns#Positive
5	http://img315.imageshack.us/img315/3918/nokia6600288ju.jpg	http://rdf.data-vocabulary.org/#rating	"5"

Figure 3 RDF Triples

Table 1 RDF to RDB mapping

Type	extractedFrom	describesFeature	hasPolarity	Rating
Opinion	http://www.amazon.co.uk/review/R1DIIVQ8WRR7OB	design	positive	5

Table 1 gives the record field "**type**," which specifies that data present in the adjacent columns provides review information about a product. This column is identified as the "name" of the table. Further, the column "**extractedFrom**" contains a URI string, which is not useful for machine learning purposes. After removing this column, the updated table is as follows:

Table 2 Opinion RDB

describesFeature	hasPolarity	Rating
design	positive	5

The context present in the review is the key to gleaning the overall sentiment of the review. At the feature level in opinion mining, "context" is understood as the clue to disambiguating the sense of the features present in a review. For instance:

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

In the above sentence, "*emphasized*" is the clue that indicates that the "chair" is a person. This is crucial information, which can be used to classify reviews clearly, based on a particular feature and its corresponding synonym group. Now, the RDB table for the review under consideration can be updated as:

Table 3 Updated opinion RDB

describesFeature	hasContext	hasPolarity	Rating
design	phone	positive	5

4. RESULTS AND DISCUSSION

In this section, the steps for extracting features and sentiments that were explained in Section 3 are exemplified via the tabulation of the feature and sentiment data obtained in this study. A comparative histogram is also given, which shows both the positive and negative sentiments within the features obtained. Here, the number of triples identified, based on extracted features, is 108, as compared to the manual extraction of features in Hu and Liu's work (2004), in which the number of triples extracted was 402. When both the triple counts are combined, there are 510 triples altogether. This count aids the categorization of the reviews.

Table 4 Feature data obtained via the chosen methodology

Frequent Features	Relevant Features	Context-dependent Features	Features that are Irrelevant (Feature Pruning)	Synonym-Group Name sets
{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}

Table 5 Sentiment data obtained via the chosen methodology for 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}

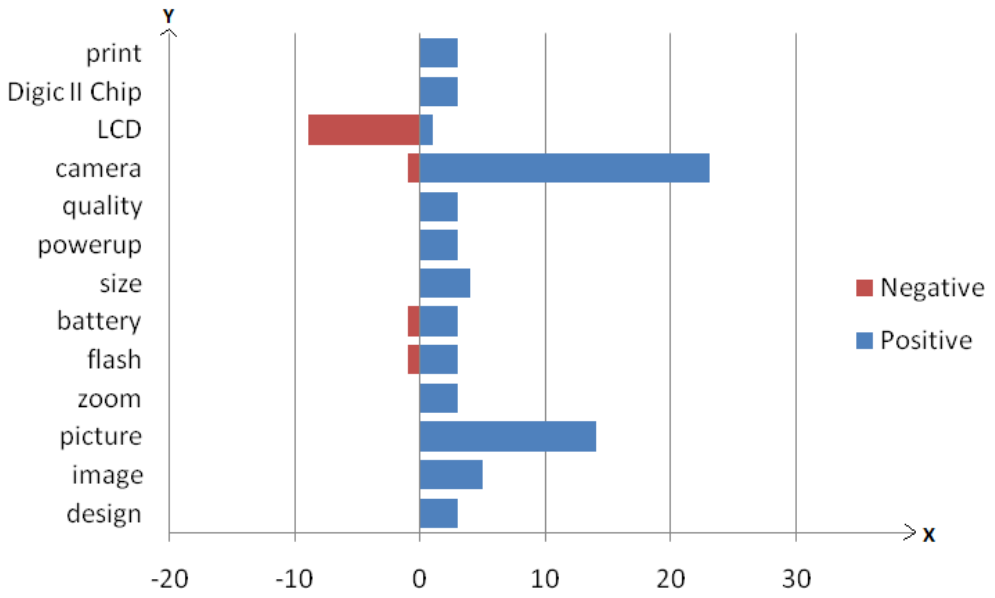

Figure 4 Comparative histogram showing the positive and negative sentiments of the features

5. CONCLUSION

The feature- and sentiment-based linked instance RDF data has been obtained successfully. The process of using the RDF database helps to classify the reviews accurately, especially when the RDFS and OWL Ontology constructs and the machine learning algorithm are also utilized. The categories identified can assist a customer in making informed decisions about purchasing products in a shorter amount of time than might be possible otherwise.

6. REFERENCES

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

Baccianella, S., Esuli, A., Sebastiani, F., 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, *LREC*, Volume 10, pp. 2200–2204

Buitelaar, P. et al., 2013. Linguistic Linked Data for Sentiment Analysis

de Freitas, L.A., Vieira, R., 2013. 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

Fellbaum, C., 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Hu, M., Liu, B., 2004. Mining and Summarizing Customer Reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 22–25, 2004, Seattle, WA, USA

Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval*, Volume 2(1–2), pp. 1–135

Peñalver-Martínez, I., Valencia-García, R., García-Sánchez, F., 2011. *Ontology-guided Approach to Feature-based Opinion Mining*. *In:* Natural Language Processing and Information Systems, Springer, Berlin and Heidelberg, Germany, pp. 193–200

Polpinij, J., Ghose, A.K., 2008. An Ontology-based Sentiment Classification Methodology for Online Consumer Reviews. In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Volume 1, pp. 518–524. IEEE Computer Society, December 2008

Softic, S., Hausenblas, M., 2008. Towards Opinion Mining through Tracing Discussions on the Web

Verma, S., Bhattacharyya, P., 2009. Incorporating Semantic Knowledge for Sentiment Analysis. In: *Proceedings of the 31st International Conference on Natural Language Processing (ICON)*, December 14–17, 2009, Hyderabad, India

Yang, C.C., Wong, Y.C., Wei, C.P., 2009. Classifying Web Review Opinions for Consumer Product Analysis. In: *Proceedings of the 11th International Conference on Electronic Commerce*, August 12–15, 2009, Taipei, Taiwan