PROO ontology development for learning feature specific sentiment relationship rules on reviews categorisation: a semantic data mining approach

D. Teja Santosh*

Computer Science and Engineering,
JNTU Kakinada,
Andhra Pradesh, India
Email: tejasantoshhd@gmail.com
*Corresponding author

B. Vishnu Vardhan

Computer Science and Engineering,
JNTUH College of Engineering Jagtial,
Telangana, India
Email: mailvishnuvardhan@gmail.com

Abstract: Crucial data like product features were obtained from consumer online reviews and sentiment words were gathered in Resource Description Format (RDF) in order to use them in meaningful reviews based categorisation on sentiments of the feature. The meaningful relationships among these pieces of RDF data are to be engineered in a Product Review Opinion Ontology (PROO). This serves as background knowledge to learn rule based sentiments expressed on product features. These semantic rules are learned on both taxonomical and non-taxonomical relations available in PROO Ontology. In order to verify the mined rules, Inductive Logic Programming (ILP) is applied on PROO. The learned ILP rules are found to be among the mined rules. The positively classified features are grouped to justify the goal of ILP for examples which are both complete and consistent. Left out negative examples are useful in knowing their count at the time of categorisation.

Keywords: PROO ontology; semantic data mining; sentiment rules; inductive logic programming; reviews categorisation.


Biographical notes: D. Teja Santosh is presently working as Assistant Professor in CSE at GITAM University Hyderabad. He is the research scholar in the faculty of Computer Science and Engineering, JNTUK, Kakinada. His areas of interest are data mining, machine learning, natural language processing. He was the Additional Reviewer for 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) in Kerala, India. He is the member of professional bodies of international repute namely IEEE, ACM, IAENG. He is one among the 16 research scholars of B. Vishnu Vardhan.

B. Vishnu Vardhan is presently working as Professor in CSE at JNTUH College of Engineering Nachupally, Kondagattu. He was an active member in Free Software Movement in India. His areas of interest are linguistic processing, data mining, natural language processing, information security and other elite fields of engineering. He has completed Government of Andhra Pradesh funded project worth 5 lakhs from Ministry of IT on localisation activity. As a co-investigator he completed another UGC funded project worth 9 lakhs. He is a technical member for International Journal of Information and Electronics Engineering.

1 Introduction

The social web has changed the perspective of understanding about the information. in this evolution, e-commerce has changed the way of providing purchase experience of consumers through online to express their sentiments on a specific product which was earlier bought. This has created the opportunity to have huge responses in terms of reviews. These reviews are preprocessed using computational linguistics (Santosh and Vardhan, 2015) for extracting features and
sentiment words. This process becomes the motivating factor for knowing the sentiment of a product on a particular feature. Many constructive recommendation systems were developed explaining the aggregate information of product with their advantages and disadvantages for systematic interpretation of the reviews and comments from the consumers appeared in blogs, reports, product websites and etc. This crucial data obtained from online reviews are structured in a Resource Description Framework (RDF) data model. The data expresses information about the features and sentiments. There is a need to provide relationships among these data items for improved understanding by the machine. Web Ontology Language (OWL) helps to achieve this. OWL compliant Product Review Opinion Ontology (PROO) is engineered on top of the RDF data layer to read the features and sentiment words automatically by the machine and also to reason the review data.

The structured RDF data is used for traditional machine learning task called classification. The automatic classification of online product reviews is not possible by providing the RDF data to the machine learning algorithm. This helps to learn the patterns inherent in the data only. It needs the domain information of product reviews as background knowledge for efficient classification. The engineered PROO Ontology serves this requirement with RDF data for precise classification. Semantic Web Rule Language (SWRL) is a high level abstract syntax used for extending the expressiveness of OWL Ontology. The SWRL rules are written to infer new relations from the developed Ontology. The engineered Ontology with SWRL rules and RDF data is mined for learning the semantic hypothesis search space using Iterative Dichotomiser 3 (ID3) (Quinlan, 1986) decision tree classification algorithm. The learned hypothesis search space is found to be with the SWRL rules.

The hypothesis search space is a collection of rules for classifying the sentiment of the review, based on the extracted features from that review. In order to verify the SWRL rules obtained from semantic hypothesis search space, the Inductive Logic Programming (ILP) rules are learned from the Ontology and RDF data using C4.5 rule learning approach. The learned ILP rules are those which best fits the RDF data from semantic hypothesis space of rules. These are found to be classifying only the positive RDF examples. Positively classified examples areclustered to justify the goal of ILP that says learned hypothesis is to be both complete and consistent. Left out negatively classified examples are useful at the time of categorising the reviews using RDF features and sentiment words.

The organisation of the paper is as follows: Motivation for carrying out this research is described in Section 2, the contributions in this direction are critically reviewed in Section 3, terminology is introduced in Section 4, the proposed method is explained in Section 5, how positively classified examples cover the maximum area under the Receiver Operating Characteristics (ROC) curve is briefly explained in Section 6, and finally, conclusions and future work are specified in Section 7.

2 Motivation

The scientific research is becoming more interactive, knowledge based and data-driven. Semantic Web Technologies such as RDF and OWL ontologies are the popular technologies offering solutions to the specific problem to handle the challenges such as semantic web services, semantic annotations etc. These technologies are changing the perspective of analysing the data by the humans.

Despite the fact that modern data mining platforms such as WEKA work with propositional data, the performance of the data mining methods are considerably enhanced by providing additional relationships among the data. Ontologies are engineered to make the machine interpret the data accurately. Serving purpose of the ontologies is used as the main information resource for mining.

Semantic Data Mining (SDM) is having a great convenience in applications where ontologies are used as background knowledge. In order to support the users of various disciplines like in sociology, finance, E-commerce, the ontologies are developed at a rapid pace. Semantic data mining algorithms have to consider the ontologies as input and semantically meaningful result as an output to the user. The PROO Ontology is developed by keeping this goal in mind.

3 Related work

Semantic Data Mining is considered as a major research work in incorporating domain knowledge for advancing various search algorithms. These search algorithms uses the semantics to realise the actual knowledge potentially hidden in the data. Liu (2010) applied the concept of semantic data mining and identified the importance of semantic annotation and proposed machine learning based semantic search algorithm for annotating semi structured data. Lavrač et al. (2011) applied semantic data mining using CN2, SEGS and g-SEGS classification algorithms on enriched gene sets searching with domain ontology as background knowledge for data mining. Mustapaša et al. (2011) applied semantic data mining by implementing Apriori Algorithm on e-learning domain by framing questionnaire to decrease the gap between itself and traditional learning. Jeon and Kim (2011) developed semantic decision tree using semantic data mining on person domain and presented rich and complex expressions but with number of limitations. Velásquez et al. (2011) worked on web personalisation on Chilean geographical information systems web application and clustered user preferences using SOFM and K-Means Algorithms by applying semantic data mining. Nebot and Berlanga (2010) applied semantic data mining on biomedical ontology to learn association rules in the biomedical data. Segura et al. (2011) used ontology as background knowledge to learn new relations among the learning objects metadata. This research analysis was carried out in the pedagogical point of view only. Abedjan and Naumann (2011) discovers new and valuable insights in the data from the semantic annotations in the RDF data by mining them at the Statement level.

Inductive Logic Programming (ILP) has taken its form in the year 1996 as machine learning in first-order language
where new relations (Knowledge Discovery) are deduced from the RDBMS deductive databases. Considerable research has undergone in this young field and still research is active in this area. Morik (1997) discovered new relations from potential customer and car engines and gearshift tables in the form of horn clauses using rule discovery tool. The author was successful in restricting (structuring) the generation of hypothesis. Lavrač et al. (2014) used the concept of ILP to learn association rules from the textual corpus by applying the Wordification methodology to find the most relevant feature from the given word.

In the process of understanding deductive rules from Ontologies and from ILP based surveys, certain shortcomings were identified: The semantic models or patterns after are learned based on the available Ontology classes and relationships by applying SDM algorithm. SWRL rules were not used as they are helpful in deducing new relations from the Ontology so that valuable insights are gained easily. These are used upon the RDBMS Ontology tables to structure, refine and restrict the models so that end user search is not compromised. Also, there are limitations in the expressed conditions which SWRL overcome. The rules generated from ILP are in restricted first order form. ILP uses first order logic whereas SWRL uses predicate logic. These ILP rules are to be generated from the hypothesis space search of SWRL rules for classifying the examples.

4 Terminology

The features introduced in this section are categorised into two ways. They are Ontology tools and machine learning topics.

4.1 Protégé Ontology Editor and Ontorion Fluent Editor

Protégé (Alani et al., 2004) is a platform for engineering Ontologies. It is open source software developed by Stanford Medical Informatics. Ontological knowledge Engineering is based on frames representation from Artificial Intelligence environment. Ontorion Fluent Editor (Kapłański, 2011) is a comprehensive tool for editing and manipulating complex ontologies. The main feature of fluent editor is the usage of Controlled English as a knowledge modelling language. The Controlled English is a subset of English language with restricted grammar and vocabulary.

4.2 Inductive logic programming

Inductive Logic Programming (ILP) (Džeroski, 1996) is viewed as machine learning in a first-order language, where relations are present in the context of deductive databases. It is relevant for knowledge discovery in relational and deductive databases, as it describes about the patterns involving in more than one relation.

4.3 Semantic data mining

Semantic Data Mining (Benites and Sapozhnikova, 2014), a data mining approach where domain ontologies are used as background knowledge for data mining. The ontology with the corresponding instance data is mapped into relational database tables which are further analysed for dependencies among these tables. This analysis leads to the creation of a new table with relational dependencies. This table is useful for generating new patterns/models.

5 Discovering concept relationship rules from PROO Ontology using semantic data mining and inductive logic programming

The principal objective of deducing/learning new relations from the PROO Ontology data by the machine is to filter the reviews based on the incoming feature request into positive and negative classes. This is carried out by retrieving product reviews written on feature hierarchy for wise decision making for a new customer wants to purchase a product. In order to achieve this goal, a framework presented in Figure 1. The framework is composed of three main modules. The first being the development of the PROO Ontology for review features and sentiments. The second module is to generate models/patterns which are analysed for new relations with SWRL rules using SDM and the last being inductive logic programming for learning new relations and verifying with the SWRL rules space. A detailed description of these components is provided below as a structured framework.

5.1 Development of the Product Review Opinion Ontology - (PROO)

In computer science and information science, ontology (Horrocks, 2009) formally represents the knowledge as a set of concepts within a domain, and the relationships between those concepts. It is used to model a domain and support reasoning about concepts. In theory, ontology is a formal, explicit specification of a shared conceptualisation. Formally, ontology is the statement of a logical theory (Gruber, 1995). Some of the reasons to engineer Ontology are namely to share common understanding of the structure of information among people or software agents and to enable reuse of domain knowledge in order to make domain assumptions that are explicit and to focus on separate domain knowledge from the operational knowledge, and lastly to analyse the domain knowledge.

The artificial intelligence literature contains many definitions of ontology and many of these contradict one another. An ontology is a formal explicit description of concepts in a domain of discourse (classes sometimes called concepts), properties of each concept describing various features and attributes of the concept (slots sometimes called roles or properties), and individual instances of classes. This constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins.

In practical terms, developing the ontology includes determining the domain and scope of Ontology, defining classes in the ontology, arranging the classes in a taxonomic (subclass–superclass) hierarchy, defining slots and describing the allowed values for these slots, filling in the values for slots for instances. PROO Ontology is engineered with the above specified steps.
5.1.1 The domain and scope of PROO ontology
The domain that the PROO Ontology covers is all the product reviews. The PROO Ontology is used to relate product parts and product part features as product features and identify the Sentiment associated with the particular feature. PROO Ontology answers the questions on product features, the object present in the review, with the feature that is with expressed sentiment, and the sentiment orientation of the review with respect to review feature.

5.1.2 Classes in the PROO ontology
The classes in the PROO Ontology are defined in the top-down approach. The creation of the general classes namely Fact, Opinion, Object, Review, Feature, Polarity, and Sentiment and so on. The specialised classes namely ObjectPart and ObjectPartFeature and TotalOpinion were created.

5.1.3 Arrangement of the classes in a taxonomic (subclass–superclass) hierarchy
The list of classes that are defined are the classes in the ontology and becomes anchors in the class hierarchy. The organisation of the classes is in a hierarchical manner. PROO Ontology class taxonomy is shown in Figure 2.

5.1.4 Slots and describing allowed values for these slots
The classes defined for PROO Ontology alone do not provide enough information to answer the competency questions. Once these classes are defined, the internal structures of concepts are to be described. These are called properties of the class. The object properties are namely expressFeature, hasObjectPart, hasObjectPartFeature, contains, isExpressedOn, isPartOf, mineFrom, hasPolarity, portrayObject etc. The Opinion slots with allowed values are shown in Figure 3.

5.1.5 The creation of values for slots with instances
The final step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires choosing a class, creating an individual instance of that class, and filling the slot values. The class Feature in Figure 4 is specified with an individual instance item and the selected value from the range class of that relationship.
The visualisation of the PROO Ontology with the classes, slots with the allowed values and the slot relationships among the classes forming a knowledge base is presented below in Figure 5.

5.2 Semantic data mining the PROO Ontology for semantic rules to learn new relations

Semantic Data Mining (Benites and Sapozhnikova, 2014), a data mining approach where domain ontologies are used as background knowledge for data mining. The formal definition of SDM for the task of classification is given below.

Formal Definition of Semantic Data Mining for Classification: Given a set of instances $I$ from a domain knowledge base $KB$, and a set of reviews $D$, find all rules of the form $A \rightarrow B$, where $A$ is disjoint combination of classes and relationships and $B$ is the class name used for learning the class label of $I$ that may occur in the reviews of $D$, and the classes $A$ and $B$ satisfy a set of constraints $C$ defined over the ontology $O$.

Mining the engineered ontology into to by the current classification algorithms is a hard task. In order to simplify the process, the ontology with the corresponding instance data is mapped into the relational database tables. These tables are further analysed for dependencies among them. This analysis leads to the creation of a new table with relational dependencies. This table is used for learning patterns/models. The analogies between ontologies and relational databases are taken into account for the semantic data mining process. These are:

- Instances (RDF data) correspond to records
- Classes correspond to tables
- Attributes correspond to record fields
- Relations correspond to relations between the tables

The slot name of a domain OWL class is considered to be the attribute and the range OWL class is filled with its corresponding instance to that slot in the mapped RDBMS table. For example the Opinion class in Ontology is an Opinion table in RDB with expressFeature, portrayObject, mineFrom and hasPolarity slots as table attributes and the values of these attributes are the instances of Feature, Object, Review and Polarity PROO Ontology classes.

Conventional data mining process learns the data model from the patterns that are inherently hidden in the data and the relations among the attributes in the dataset. Semantic Data Mining, in contrast to conventional data mining mines the abundance of patterns available in the ontology and learns the semantic associations among the classes from the ontology. The mining process is carried out by searching for those predicates which are available in PROO Ontology that satisfy taxonomical and non-taxonomical constraints on the ontology at the time of hypothesis rules space learning.
In order to carry out this search process, content constraints are defined over the classes (tables) and relationships (relations) that are available in the ontology. A content constraint is a rule that is able to capture other knowledge from the ontology in addition to the available abundant patterns. A content constraint is either a taxonomical constraint or a non-taxonomical constraint. A taxonomical constraint is a content constraint expressed on the predicate written between the classes that are organised in the hierarchical manner in the ontology. A non-taxonomical constraint is a content constraint expressed on the predicate written between the related classes. This form of learning semantic association rules in the form of relationships among the data instances was specified by Antunes (2007) for efficient Association rule mining. The detailed procedures expressed in algorithmic form for learning taxonomical constraints and non-taxonomical constraints are presented below.

Algorithm for learning taxonomical constraints from Ontology:

```
Input: Ontology {O}
Output: semantic rule {A → B}
LEARN_TAXONOMICAL_CONSTRAINT (Ontology O)
{
  for each node in the taxonomy from Ontology O
  {
    content_constraint = false;
    if(parentof(parentnode, childnode))
      content_constraint=true;
      write (parentof(parentnode, childnode) → targetclass(object));
    else if(parentof(parentnode1, childnode1) ^
            parentof(parentnode2, childnode2))
      childnode1 ← parentnode2;
      content_constraint=true;
      write(parentof(parentnode1,childnode2) →
            targetclass(object));
  }
}
```

Algorithm for learning non-taxonomical constraints from Ontology:

```
Input: Ontology {O}
Output: semantic rule {A → B}
LEARN_NONTAXONOMICAL_CONSTRAINT (Ontology O)
{
  for each node in Ontology O
  {
    content_constraint = false;
    if(objectproperty(nodei, nodej))
      content_constraint=true;
      write (objectproperty(nodei, nodej) → targetclass(object));
    else if(objectproperty(nodei, nodej) ^
            datatypeproperty(nodej, rel(int)))
      content_constraint=true;
      write(objectproperty(nodei,nodej) ^
            datatypeproperty(nodej, rel(int)) → targetclass(object));
  }
}
```

Algorithm for learning taxonomical constraints works as follows: given the ontology, all the super class and sub class hierarchies are identified. Super class node is called parent and sub class node is called child. The rules are then generated with the predicate on the hierarchy as the relation between parent and child leading to target class. Content constraint is initialised to the value false in the beginning and is changed to true value, once related class nodes are obtained. The rules are then generated with the object property as the relation between the related classes leading to the target class. The Algorithm also checks for the relation between related classes and datatype properties. The related class node and the datatype property are related using the conditions that are imposed on the ontology is identified by the algorithm. The content constraint value is changed to true. The rules are then generated with the relation between object property and the datatype property leading to the target class.

As explained earlier out of the six rules the first four rules are taxonomical constraints and remaining two rules are non-taxonomical constraints. These rules are specified separately in Appendix A.

The Opinion, Feature, ObjectPart and ObjectPartFeature tables are used in generating the semantic rules on Sentiment table which says that the corresponding object part and object part feature are the features that possess positively oriented sentiment when the rating on these features has a value of 3. The corresponding class hierarchies and the related classes of the PROO Ontology are presented in Figure 6.
Humans trust a review by carrying out a basic statistical analysis of finding the percentage of people expressing positive opinion on a particular feature. In order for the machine to carry out the same statistical analysis, confidence intervals are to be found out. The confidence that the opinion is expressed on the feature is carried out by calculating the confidence intervals of the likely opinion expressed people. The obtained range specifies with confidence that X% people have expressed positive opinion on the feature with rating of the feature taken into consideration and Y% people have expressed against the feature with rating of the feature taken into consideration. The formula to calculate the confidence intervals is:

\[ p \pm (CI) \times \sqrt{\frac{p(1-p)}{n}} \]  
(1)

The value of \( p \) is the percentage of people has positive opinion for a feature with rating of the feature taken into consideration, \( n \) is the sample size and \( CI \) is Confidence Interval (Mitchell, 1997) has constant 1.96 for 95% confidence and 2.58 for 99% confidence. To test whether the newly added property influences the rating property, an independent \( t \)-test and to test whether the newly added property affects the Sentiment class, Binary Logistic Regression analysis is carried out. The statistics are presented in Figure 7 below.

The statistical analysis revealed that rating property influences the Sentiment class in categorising the reviews and also the newly added property “hasStatisticalTrust” do not influence the rating attribute and does not change the data model. This attribute do not influence the Sentiment class. The “hasStatisticalTrust” property is useful for making the machine learn the numerical trust data and provide information at the time of review categorisation. The learned model in the form of Decision Tree from the sample RDF dataset is tabulated in Table 1.

### Table 1  
<table>
<thead>
<tr>
<th>Decision Tree Model learning in Sipina</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning inputs</td>
</tr>
<tr>
<td>Classifier learned</td>
</tr>
<tr>
<td>Splitting Attribute</td>
</tr>
<tr>
<td>Number of examples classified under Good Sentiment class label</td>
</tr>
<tr>
<td>True Positives Percentage</td>
</tr>
<tr>
<td>Application of Rule Selection</td>
</tr>
</tbody>
</table>

5.3 Inductive logic programming for verification of the semantic hypothesis search space

Given some positive examples and negative examples, Inductive Logic Programming finds/learns rule by using PROO Ontology as the background knowledge (understood as Hypothesis Space of SWRL rules to be searched) such that the rule covers all the positive examples and the rule does not cover any negative examples. The rule satisfying the mentioned first condition is said to be complete and the rule satisfying the mentioned second condition is said to be consistent. The learned conjunctive rules from the PROO Ontology using ILP’s restricted first order logic are specified in Appendix A. The learned model in the form of C4.5 rules from the sample RDF dataset is tabulated in Table 2.

These ILP rules learned are considered to follow Occam’s razor principle (Hamilton, 1862) of finding the hypothesis that best fits (the disjunctive expression in the rule antecedent provide simplest possible explanation for learning the rule consequent) the RDF data as these are with the PROO
Ontology SWRL rules space. These ILP rules have intuitive inferencing capabilities so as the SWRL rules. Both of these rule forms are found to learn the positive sentiment class label for the positive examples.

### Table 2  Rule learning in Sipina

<table>
<thead>
<tr>
<th>Machine Learning inputs</th>
<th>Values considered/generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier learned</td>
<td>C4.5 Rules</td>
</tr>
<tr>
<td>Splitting Attribute</td>
<td>Rating</td>
</tr>
<tr>
<td>Number of examples</td>
<td>14</td>
</tr>
<tr>
<td>True Positives Percentage</td>
<td>93%</td>
</tr>
<tr>
<td>Application of Rule Selection</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### 6 Experimental results and discussion

The data corpus that is used in the feature specific sentiment classification is the collection of electronic device reviews from Amazon. The electronic device being Nokia 6610 cellular phone. The selection of the reviews was taken in such a way that each review contains explicit specification of product features. The number of selected reviews was 4000, out of which 2000 reviews are positive and the rest are negative.

The pre-processing of data is carried out by removing stop words and non English words. Negation words were present near to the adjective in a review sentence and those are handled with care. For those review sentences, the sentiment orientation of the word is captured by simply flipping the actual sentiment. We have extracted features by using the Natural Language Processing approach specified earlier in (Santosh and Vardhan, 2015). The number of extracted features were 42 whereas when we compare with Hu and Liu’s (2004) work the selected features were 67. The fact that adjectives written adjacent to the features in the review will reflect the experience of the consumer while specifying and these adjectives are considered to be the sentiment words. The positive and negative sentiments of these sentiment words of the extracted features are calculated using SentiWordNet (Baccianella et al., 2010). The obtained features are sentiments that are structured in RDF data model. This structured dataset is given as an instance data to the PROO Ontology classes. The slots are used to relate the class instances from the RDF data. We have considered only one product type for the analysis as the PROO Ontology is developed for a class of mobile phones of different manufacturers.

ILP rules are learned from the PROO Ontology. The rule antecedent is learned by forming a conjunction of PROO Ontology classes and the relevant slots which relate to these classes. The class instances and the slot values are reasoned for learning the target class instance which is the rule consequent. The obtained rules cover the positive examples. The evaluation of the learned rules was visualised with the help of Area Under Receiver Operating Characteristic Curve (AUC) presented in Figure 8. The AUC is a measure to showcase the examples covered in either of the two sentiment groups (good/bad) available from the dataset.

### Figure 8  Area under ROC Plot with classifier accuracy

The parameters of the Receiver Operating Characteristic (ROC) curve are the target class label and the ranking attribute. The target class label considered is good for the Sentiment class and the ranking attribute is considered as rating. From the dataset containing 2000 positive reviews, the ILP rules correctly covered 1833 positive reviews. This resulted in an accuracy of 91.67% of ROC area coverage. This signifies that the ILP rules learned are able to clearly categorise the reviews for a particular feature when the customer does the selection of that feature.

The parameters provided in Table 3 below compares the area covered under ROC curve between the works done by Fang and Zhang (2015) with the results obtained from the current work presented in Figure 8. They used Naïve Bayesian Classifier on the product features data for the task of sentiment analysis. The performance of the Naïve Bayesian classifier was purely dependent on the complete training data. No ontology was used. The classifier performance is evaluated for categorising the reviews at sentence level. In our work, the classifier used is the collection of semantic rules obtained by mining the ontology. These rules are found to categorise the reviews at feature level which is better than categorising the reviews at sentence level. The classifier outputs the positive sentiment class label which provides clear information about the cluster of positive reviews when the search takes place on a particular feature.

### Table 3  Classifier performance parameters comparison

<table>
<thead>
<tr>
<th>Classifier used</th>
<th>Ontology used</th>
<th>Review Categorisation level</th>
<th>Review Rating format</th>
<th>ROC Step start</th>
<th>ROC Step End</th>
<th>ROC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayesian (Fang and Zhang, 2015)</td>
<td>No</td>
<td>Sentence level</td>
<td>star scale</td>
<td>0.6</td>
<td>0.9</td>
<td>90%</td>
</tr>
<tr>
<td>SDM+ILP (Our work)</td>
<td>Yes</td>
<td>Feature level</td>
<td>number scale</td>
<td>0.75</td>
<td>0.91</td>
<td>91.67%</td>
</tr>
</tbody>
</table>
It is found that there is an increase of 1.67% in the accuracy of the classifier achieved in our work. This signifies that mining Ontology with the data helps in improving the learning of a machine for automatic categorisation of online reviews than mining the data only.

7 Conclusion and future work

The discovery of sentiment relationship rules from PROO Ontology using semantic data mining and Inductive Logic Programming was carried out successfully. Algorithms for learning taxonomical and non-taxonomical constraints were presented for mining the ontology leading to the generation of more general rules. In the process of this work it is found that the newly learned rules are able to automate the process of review categorisation on the basis of selected features by the machine. A new attribute named “Statistical Trust” was introduced and was tested for affecting the target Sentiment class. The generated ROC curve was compared against the Naïve Bayesian classifier to analyse the performance of the ontology supported classifier. The ontology supported classifier is found to be better than the Naïve Bayesian classifier in terms of performance.

In future, the PROO ontology is extended to include different classes of mobile phones available from different manufacturers. This helps in comparison of different classes of phones for comprehensive set of features in the semantic manner.

References

Appendix A

A.1 SWRL rules of PROO Ontology

The SWRL rules written to improve the expressiveness of the PROO Ontology are the same as the taxonomical and non-taxonomical constraints applied on the ontology. The taxonomical and non-taxonomical constraints based on the algorithms specified in section 5.2 that are learned from PROO Ontology which are found to be SWRL rules are given below.

- `isObjectPartFeatureOf(?z, ?y) ∧ rating(?y, 3) → Sentiment(?z)`
- `isObjectPartOf(?x, ?y) ∧ rating(?y, 3) → Sentiment(?x)`
- `isObjectPartFeatureOf(?z, ?y) ∧ Sentiment(?y) → Sentiment(?z)`
- `isObjectPartOf(?x, ?y) ∧ Sentiment(?y) → Sentiment(?x)`
- `expressFeature(?p, ?y) → Sentiment(?p)`
- `hasPolarity(?p, ?q) → Sentiment(?p)`

The Reasoner of Ontorion Fluent Editor helps to query the Ontology in the Controlled Natural Language with its corresponding Grammar for understanding classes, properties and instances. The SWRL rules are executed on the Ontology constructs. The SWRL reasoned conclusions from Ontorion Fluent Editor is illustrated in Figure A1 and is given below.

A.2 ILP rules from PROO Ontology

The ILP rules learned based on the deductive database of relational tables specified in Table 2 are given below.

- `isObjectPartOf(ObjectPart, Feature) & Sentiment_Good(Feature) → Sentiment_Good(ObjectPart)`
- `isObjectPartFeatureOf(ObjectPartFeature, Feature) & Sentiment_Good(Feature) → Sentiment_Good(ObjectPartFeature)`
- `isObjectPartOf(ObjectPart, Feature) & rating(Feature, >=3) → Sentiment_Good(ObjectPart)`
- `isObjectPartFeatureOf(ObjectPartFeature, Feature) & rating(Feature, >=3) → Sentiment_Good(ObjectPartFeature)`

Out of the four ILP rules, the 3rd and 4th rules are the SWRL rules and also are the semantic data mining learned hypotheses. These rules are considered as the hypotheses for learning appropriate Sentiment class label.

Figure A1 Debugging SWRL rule for Sentiment expressed on Feature in Fluent Editor