Ontology based combined approach for Sentiment Classification

Khin Phyu Phyu Shein

***Abstract*—**Text documents contain opinions or sentiments about the objects, such as movie reviews, book reviews, and product reviews etc. Sentiment analysis is the mining the sentiment or opinion words and identification and analysis of the opinion and arguments in the text. In this paper, we proposed an ontology based combination approach to enhance the exits approaches of sentiment classifications and use supervised learning techniques for classifications.

***Keywords*—** ontology using formal concept analysis design, opinion mining, part of speech tagging, sentiment analysis, support vector machine.

I. Introduction

There are two main categories in textual information, they are facts and opinions. Facts are the objective statements and there is a lot of research on the information retrieval. Opinions are the subjective statements and still rare in existing researches. Opinions reflect the people's sentiments or opinions about the product and events.

Many of the existing research based on mining and retrieval of factual information not on opinions. Opinions are also important when someone wants to hear the other's opinions before they make a decision.

There are two types of opinions:

Direct opinion- opinion expressing on the products, events, topics, persons, etc.

Eg. "AmityVille Horror"-Not a great movie, but Ryan Reynolds yet again proves how great of an actor that he is. Somewhat campy, but not very original.

Manuscript received October 9, 2001. (Write the date on which you submitted your paper for review.) This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

Comparisons- express the similarities or differences between more than one object.

Eg. Scenes in "A Beautiful Mind" is better than "No Country for Old Men"

Opinion mining extract user's opinion on products, political issues, etc., which are express in online forums, blogs or comments. And then find that the opinions are positive or negative on a particular object or some features of the object and this classification is also called the sentiment analysis.

The result can be used in business and organization by consultants, surveys and focused groups, etc., individuals who are interested in other's opinions for purchasing a product or using a service, finding opinions on political topics, in advertising placements and in opinion retrieval/search.

Most of the existing research based on the product review. But in this paper we focus on the movie review domain because this domain is different from the other product reviews. When a person write a movie review, he probably comments not only movie elements (eg., romance, musical, script) but also on the movie related people (eg., director, composer, actors). Therefore the commented features in the movie review are more complex than the products review and have more challenging. So we choose to use the movie review domain for our combination approach for sentiment classification.

In this paper, we describe the sentiment classification based on the combination approach of Natural Language Processing (NLP), Formal Concept Analysis (FCA) based ontology on movie review domain and Support Vector Machine (SVM) classifier.

The remainder of this paper is organized as follows. Section 2 describes some related work. Section 3 states the motivation of the proposed approach. Section 4 introduced the proposed approach. In section 5 experimental results are provided. Finally, the conclusion and future work are presented in section 6.

II. Related Work

Several techniques are used for the opinion mining tasks. To extract opinions, machine learning method and lexical pattern extraction methods are used by many researchers [1]. In 2002, Turney [10] introduced the results of review classification by considering the algebraic sum of the orientation of terms as respective of the orientation of the documents but more

sophisticated approaches are introduced by focus on some specific tasks such as finding the sentiment of words by Hatzivassiloglou [5], Wibe [12], Riloff et al [2], Whitelaw et al [11], Dave et al [6], subjective expression by Wilson et al [13]. They use the data of review from automobiles, bank, movies, and travel destinations. They classified words into two classes (positive or negative) and counts on overall positive or negative score for the text. If the documents contain more positive than negative terms, it is assumed as positive document otherwise it is negative. These classifications are based on document and sentence level classification. These classifications are useful and improve the effectiveness of a sentiment classification but cannot find what the opinion holder liked and disliked on each feature. So they are not always true in some cases.

Pang et al [9], Mukras R.J [8] and Michael G [3] use the data of movie review, customer feedback review and product review. They use the several statistical feature selection methods and directly apply the machine learning techniques. These experiments show that machine learning techniques only is not well performing on sentiment classification. They show that the present or absent of word seems to be more indicative of the content rather than the frequency for a word.

III. MOTIVATION

Existing techniques that involve checking the similarity between a text and the seed list of words is not sufficient. Therefore we proposed the combination approach for the efficient sentiment classification. In this approach we used domain ontology and use supervised learning technique to extract the features and opinions from the movie reviews or comments to enhance the existing sentiment classification tasks. We identify the direct opinions on the feature level to get the opinions of each feature.

As our contribution we developed the domain ontology of the problem domain based on the FCA design, [7], [4]. And then use the SVM classifier to train the features based on the supervised learning which classified the opinions or sentiments in the reviews or comments as positive or negative on each feature.

IV. PROPOSED SYSTEM

In this system we focus on the feature level sentiment classification. There are three main parts in our approach: assigning the POS tags, identifying domain related features and classifying the sentiment words. We use Part Of Speech (POS) tagger to assign POS tags to words in a sentence (such as: tags for nouns, verbs, and adjective). After that we use domain ontology to extract the related concepts and attributes and then use Support Vector Machine (SVM) classifier for labeling positive or negative on the related concepts and attributes.

Our proposed system architecture is as shown in the following figure.

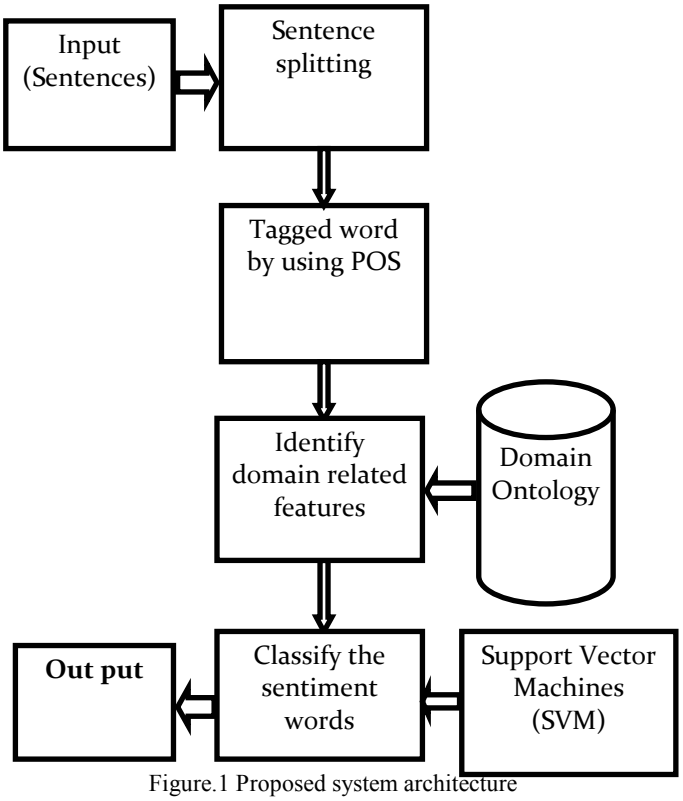

Figure.1 Proposed system architecture

A. Assigning the POS tags

To implement this process, firstly we need to do sentence splitting. It generally consider '.', '!' and '?' as sentence delimiters for splitting process, although allowances are given for the occurrence of '.' in abbreviations, number, URLs etc. For identifying sentiment phrases we use POS tagging. To implement this process we use Brill Tagger. This tagger is based on transformation based error driven learning, a technique that has been effective in a number of natural language applications which include part of speech and word sense tagging, prepositional phrase attachment, and syntactic parsing. The POS tagger assigns POS tags to words in a sentence (such as tags for nouns, verbs, and adjective). These tags are used to extract the certain features related to the proposed movie review domain.

Eg. HOOSIERS is the first sports movie since THE NATURAL that I've liked, and maybe that's not surprising.

HOOSIERS/NNP is/VBZ the/DT first/JJ sports/NNS movie/NN since/IN THE/DT NATURAL/NN that/IN I/PRP 've/VBP liked/VBN ,/, and/CC maybe/RB that/DT 's/VBZ not/RB surprising/JJ ./.

B. Identifying the domain related features

We use the domain ontology to get the domain related features. Ontology aims to provide knowledge about specific domain that are understandable by both developers and

computers and necessary for knowledge representation and knowledge exchange. Using existing taxonomical hierarchies are not enough for knowledge exchange or for informational retrieval. By using the taxonomical ordering, the concepts have no other differentiating attributes. It is not easy to change the frames and their slots once they are defined. These may cause the problems in knowledge sharing. Therefore we need a better way to describe the concepts and relation.

We present a method that is based on Formal Concept Analysis (FCA), used for analyzing data and forming semantic structures that are formal abstraction of concepts of human thoughts and identify conceptual structures among data sets. It also allows the analysis of complex structures and the discovery of dependencies within the data. In FCA, the elements of one type are called "formal objects", the elements of the other type are called "formal attributes". The adjective "formal" is used to emphasize that these are formal notions. "Formal objects" need not be "objects" in any kind of common sense meaning of "object". But the use of "objects" and "attributes" is indicative because in many applications it may be useful to choose object-like items as formal concepts and to choose their features or characteristics as formal attributes.

The main characteristics of FCA

- concepts are described by properties
- the properties determine the hierarchy of the concepts
- when the properties of different concepts are the same, then the concepts are the same

Contexts in FCA are triples (O,A,R)

Where O=finite set of object

A=finite set of attributes

R=binary relation on O and A

The procedure of designing ontology supported by a tool that use FCA is described in the Figure 2.

1. Start with empty set of concepts and properties
2. Add concepts and properties
3. Modify the ontology by the following ways
   a. Direct editing
      i. Add or remove concept
      ii. Add or remove property
      iii. Assign a property to concept or remove a property from the concept
   b. Editing as suggested by the ontology design tool
      i. When two concepts fall into one place, merge or add a distinction
      ii. Can generate the concepts which are formed by properties and super concepts of defined concepts, that are not explicitly mentioned in the concept table
4. Repeat until complete the ontology

Figure.2 Outline of the algorithm for designing ontology using FCA

TABLE 1. EXAMPLE ONTOLOGY DESIGN WITH FCA

	Good	Poor	Well done	Mindless	Reveal	Tedious
Main character	X		X		X	
plot		X				X
senses		X		X		X
action	X		X			
character		X		X		X

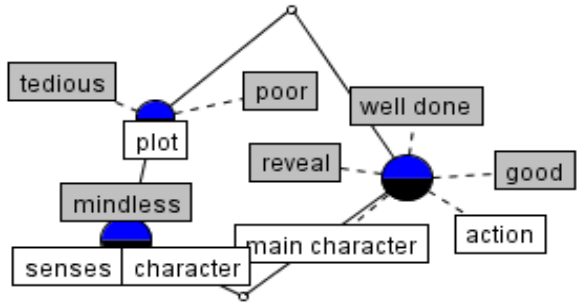

Figure.3 Concept lattice form of some features of movie review domain

We developed the domain ontology in OWL (Web Ontology Language) based on FCA design. Implementation of domain ontology is used protégé 2000. Features that are parts of the domain are extracted to be classified by SVM.

C. *Classifying the sentiment words*

We use linear SVM for feature level sentiment classification. Features are classified by linearly separated hyper plane with the binary classification. Then define the features by labelling positive or negative. We use the following equation for classification.

A binary SVM is a maximum margin classifier. Given a set of training data {x1,x2,…,xk}, with corresponding labels y1,y2,…,yk ϵ{+1, -1}, a binary SVM divides the inputs space into two regions at a decision boundary, which is separating hyperplane (w,x)+b=0 (Figure 5). If we scale w and b to make the closet point(s) to the hyperplane satisfy |hw, xii + b| = 1, then the margin equals 1/||w|| and the problem can be formulated as:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

Subject to, $x_i.w+ b >0$ for $y_i=+1$
$x_i.w+ b <0$ for $y_i=-1$, $i=1,2,\ldots,D$

The result is a hyper-plane that has the largest distance to xi from both sides. The classification task can then be formulated as discovering which side of the hyper-plane of a test sample falls into, which means that for yi=+1, $x_i.w+ b >0$ then xi falls into the positive sentiment, and for yi=-1, $x_i.w+ b <0$ then xi falls into the negative sentiment.

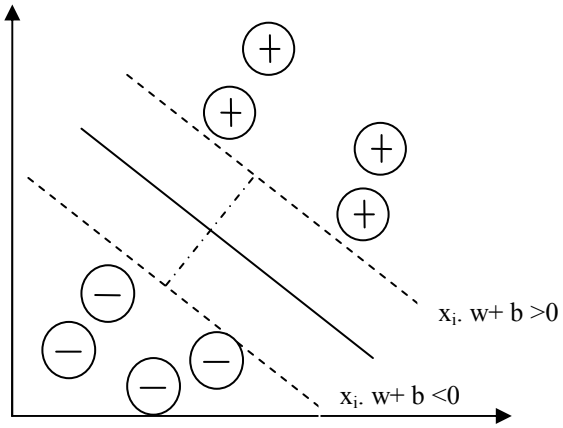

Figure.4 An illustration of the SVM method

V. EXPERIMENTAL RESULTS

To evaluate sentiment classification system, we use the customer review of a few movies form IMDB corpus which were grouped into positive and negative categories for content analysis and tested and compared with the manually tagged set of around 300 movie reviews. The results and comparisons are shown in the following table Table 2.

TABLE 2. RESULTS AND COMPARISONS OF MOVIE REVIEWS

	Without the use of domain ontology	Proposed approach
Positive sentences	Accuracy-72% Recall-90%	Accuracy-80% Recall-90%
Negative sentence	Accuracy-64% Recall-85%	Accuracy-76% Recall-85%
Ambiguous sentences	Accuracy-65% Recall-90%	Accuracy-80%% Recall-90%
Neutral sentences	Accuracy-60% Recall-90%	Accuracy-75% Recall-90%

VI. CONCLUSION AND FUTURE WORK

We proposed the combination approach of POS tagging, FCA-based domain ontology and SVM classifier intend to enhance the sentiment classification. By using this approach we can view the strength or weakness of the products or objects more detail and we hope will be useful for further development and improvement of the development and improvement of the products or objects. As the future work we need to tested with a large amount of data sets and require further training and classification to solve the problem of the comparative sentences.

REFERENCES

[1] Bo Pang and Lillian Lee. Using very simple statistics for review search: An exploration. In Proceedings of the International Conference on Computational Linguistics (COLING), 2008. Poster paper.
[2] Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing.
[3] Gamon Michael. Sentiment classification on customer feedback data, 2004
[4] Genter B, Wille, R. Formal Concept Analysis, Mathematical Foundation, Berlin, Springer Verlag 1999.
[5] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.
[6] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, pages 519–528, 2003.
[7] Marek Obitko et al. Ontology Design with Formal Concept Analysis. In CLA 2004 , pp,111-119,ISBN 80-248-0597-9
[8] Mukras R, Carroll J. A comparison of machine learning techniques applied to sentiment classification, 2004.
[9] Pang et al, Sentiment Classification using machine learning methods. In EMNLP-2002.
[10] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the Association for Computational Linguistics (ACL), pages417–424, 2002.
[11] Robert Tumarkin and Robert F. Whitelaw. News or noise? Internet postings and stock prices. Financial Analysts Journal, 57(3):41–51, May/June 2001.
[12] Wiebe J.M., Learning subjective adjective from corpora, AAAI-2000, 2000.
[13] Wilson T, Wiebe J, and Hwa R. Just how mad are you? Finding strong and weak opinion clauses. AAAI-2004, 2004.