FITM: FEROM Informal Text Mining

Surekha R. Janrao [#1], Dr. Lata Ragha [#2]

[#] Dept. Computer Engineering, Terna Engineering College
Nerul, Navi Mumbai-400706, MS, India.

[1] surekh_99@yahoo.co.in
[2] lata.ragha@gmail.com

***Abstract*— Opinion Mining refers to the broad area of natural language processing, computational linguistics and text mining involving the computational study of opinions, sentiments and emotions expressed in text. A thought, view, or attitude based on emotion instead of reason is often referred to as a sentiment. Sentiment classification is an important data mining task. Previous researches tried various machine learning techniques while didn't make full use of the difference among features. However, previous research on feature based opinion mining has not had good results due to drawbacks, such as selecting a feature considering only syntactical grammar information or treating features with similar meanings as different. To solve these problems, we propose an enhanced feature extraction and refinement method for Informal Text mining called FITM that effectively extracts correct features from review data and also improves the performance of sentiment analysis classification. This system also works for Informal Text Reviews.**

***Keywords*-NLP, Parser, FEROM, feature-based Opinion mining, feature extraction, feature refinement**

I. INTRODUCTION

Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The origin and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro blogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. Sentiment analysis systems have found their applications in almost every business and social domain [1].

A Sentiment Analysis Classification Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Much research exists on sentiment analysis of user opinion data, which mainly judges the polarities of user reviews. In these studies, sentiment analysis is often conducted at one of the three levels: the document level, sentence level, or feature level. In relation to sentiment analysis, the literature survey done indicates two types of techniques including machine learning and semantic orientation. In addition to that, the nature language processing techniques (NLP) is used in this area, especially in the document sentiment detection. Current-day sentiment detection is thus a discipline at the crossroads of NLP and Information retrieval, and as such it shares a number of characteristics with other tasks such as information extraction and text-mining, computational linguistics, psychology and predicative analysis [2].

Sentiment analysis is language processing task that uses a computational approach to identify opinionated content and categorize it as positive or negative. The unstructured textual data on the Web often carries expression of opinions of users. Sentiment analysis tries to identify the expressions of opinion and mood of writers. A simple sentiment analysis algorithm attempts to classify a document as 'positive' or 'negative', based on the opinion expressed in it.

A sentiment analysis algorithm classifies each document d Є D into one of the two classes, positive and negative. Positive label denotes that the document d expresses a positive opinion and negative label means that d expresses a negative opinion of the user. More sophisticated algorithms try to identify the sentiment at sentence-level, feature-level or entity-level [2].

B. Levels of Sentiment Analysis Sentiment analysis can be broadly classified based on the level at which it is done with the various levels being namely,

- The document level,
- The sentence level
- The feature level.

At the document level, sentiment classification of documents into positive, negative, and neutral polarities is done with the assumption made that each document focuses on a single object O (although this is not necessarily the case in many realistic situations such as discussion forum posts) and contains opinion from a single opinion holder.

At the sentence level, identification of subjective or opinionated sentences amongst the corpus is done by classifying data into objective (lack of opinion) and subjective

or opinionated text. Subsequently, sentiment classification of the above mentioned sentences is done moving each sentence into positive, negative and neutral classes.

At the feature level, the various tasks that are looked at are:

- Task1: Identifying and extracting object features that have been commented on in each review/text.
- Task 2: Determining whether the opinions on the features are positive, negative or neutral.
- Task 3: Grouping feature synonyms and producing a feature-based opinion summary of multiple reviews/text.

When both F (the set of features) and W (synonym of each feature) are unknown, all three tasks need to be performed. If F is known but W is unknown, all three tasks are needed, but Task 3 is easier. It narrows down to the problem of matching discovered features with the set of given features F. When both W and F are known, only task 2 is needed.

C. Disadvantages in Document level and Sentence level Sentiment analysis

- It does not give details on what people liked or disliked.
- Specific features of an object that the author likes or dislikes cannot be identified.
- It is not easily applicable to non-reviews, e.g. forum and blog postings.

Main focus may not be evaluation or review, but still contain a few opinion sentences.

II. Feature Based Sentiment Analysis

Sentiment classifications at both document and sentence (or clause) level are useful, however they do not find what the opinion holder liked and disliked. A negative sentiment on an object does not mean that the opinion holder dislikes everything about the object. Similarly, a positive sentiment on an object does not mean that the opinion holder likes everything about the object. Thus, sentiment analysis at the feature level is necessary.

Considering user reviews as our primary source of opinionated data in this model, we look at different review formats as shown in Figure1 that the system is expected to handle (most of which are commonly used in websites and forums) [3].

- Format 1 - Pros, Cons and detailed review: The reviewer is asked to describe Pros and Cons separately and also write a detailed review. Epinions.com uses this format.
- Format 2 - Pros and Cons: The reviewer is asked to describe Pros and Cons separately. C|net.com used to use this format.
- Format 3 - free format: The reviewer can write freely, i.e., no separation of Pros and Cons. Amazon.com uses this format.

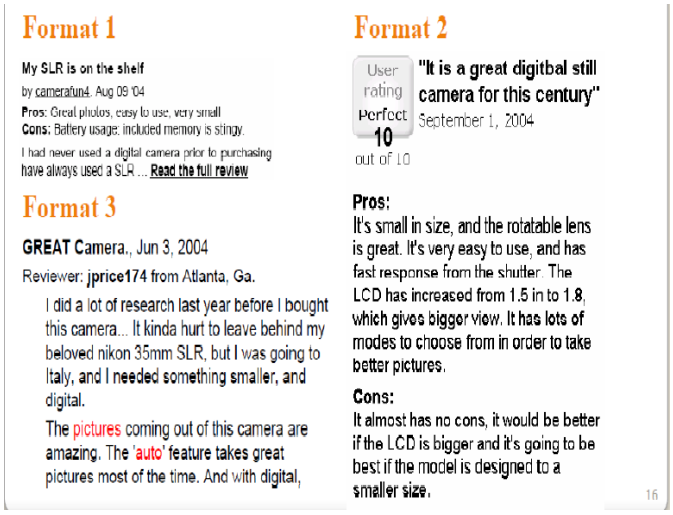

Figure 1 Different Review Formats

A. System Structure

The following system structure for feature based sentiment analysis aims to identify semantic orientations of movie reviews. This task can be conducted in eight main steps as shown in Figure 2.

- The first step is to parse the movie reviews by Stanford parser.
- The second step is to extract adjectives and features from reviews with their grammatical relationships, i.e. to take grammar analysis of reviews. In this step, all adjectives and the features modified or described by them are found. This step is conducted with grammatical knowledge.
- This step is to predict the polarity of adjectives. Based on WordNet, all adjectives are divided into five groups. The polarity knowledge library of adjectives is output in this step.
- The similarity between features and topic is computed.
- Based on polarity library and similarity between features and topic, the value of original score of every noun is produced.
- This step is to produce the value of weighted score.
- In order to predict the semantic orientation of reviews, it must be represented with a vector. This step is to construct the vector of review with weighted score.
- The last step is to produce polarity labels for reviews. We use two methods machine learning technology and total weighted score computing to

generate polarity labels of reviews and compare their precision with simple term counting method.

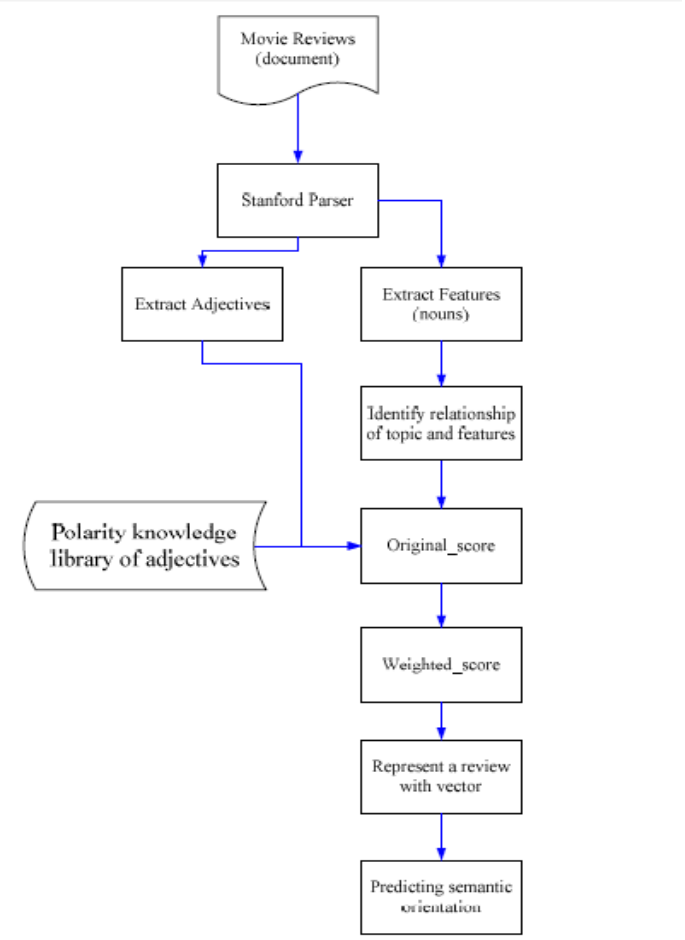

Figure 2. System Structure for feature Based Sentiment Analysis

III. Proposed Work for Improving Sentiment Analysis Classification

Previous studies on feature-based opinion mining have applied various methods for feature extraction and refinement, including NLP and statistical methods. However, these analyses revealed two main problems. First, most systems select the feature from a sentence by considering only information about the term itself, for example, term frequency, not bothering to consider the relationship between the term and the related opinion phrases in the sentence. As a result, there is a high probability that the wrong terms will be chosen as features. Second, words like 'photo,' 'picture,' and 'image' that have the same or similar meanings are treated as different features since most methods only employ surface or grammatical analysis for feature differentiation. This results in the extraction of too many features from the review data, often causing incorrect opinion analysis and providing an inappropriate summary of the review analysis. To resolve these problems, we proposes an enhanced method called, feature extraction and refinement for Informal Text Mining (FITM). The purpose of the analysis is to extract, organize, and classify the information contained in the required documents. The proposed method is based on object- oriented approach to software development. In this section, we present the architecture and functional detail of the proposed opinion mining system to identify feature-opinion pairs and Informal Text Mining also. Figure 3 presents the complete architecture of the proposed opinion mining system that is FITM (FEROM Informal Text Mining), which consists of different functional components.

A. FITM System Architecture:

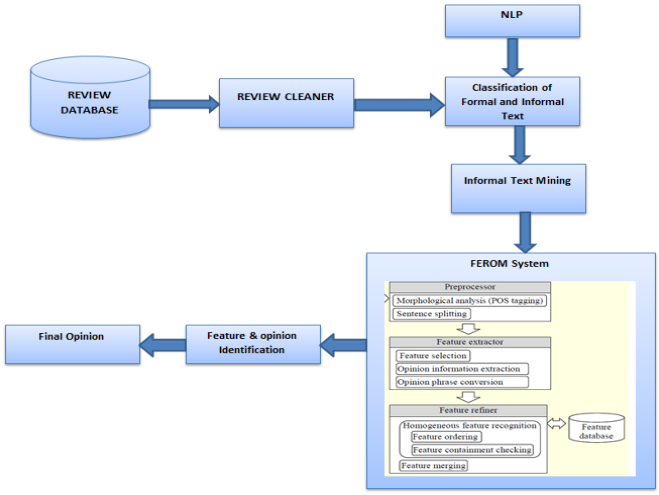

Figure 3. FITM System Architecture

The overall process of FITM consists of four phases: Review Extraction, Classification of Formal and Informal text, Informal Text Mining, FEROM system phase which has been explained in the following section.

*1) Review Extraction:*This phase consists of Review Document Extraction and Review Cleaner functional components.

- **Review Document Extraction**
  In this module the crawler retrieves reviews document from sources such as web. Then Locate and download the reviews.
- **Review Cleaning**
  After that review document is processed to review cleaning or filtering. Filtering process, filter out or remove noisy review.

2) Classification of Formal and Informal Text: After removing noisy review classify the remaining data review according to formal and informal style [4]. Filtered review document are divided into manageable record size chunk. This

is assign as input for document pre-processor to Parts of Speech tag (POS) to each word, like Stanford Parser [5]. It converts each sentence into set of dependency relationship between pair of words.

3) Informal Text Mining: Following steps for Informal Text Mining system: where, Input = Informal review sentences

1. Identification of the Informal sentence reviews.
2. Add it to data pre-processor.
3. Apply Parts of speech tagging.
4. Identification of noun, verb, adverb.
5. Apply rules as in paper [6] if reviews are not informal.

Large number of noun, verb, adjective are extracted which gives features and opinion represented as undirected graph as shown in Figure 4.

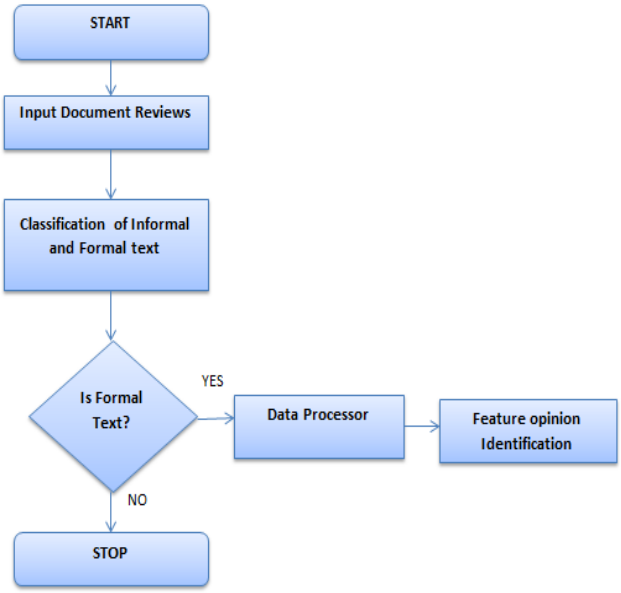

Figure 4. Flow chart for Informal Text Mining

4) FEROM (Feature Extraction and Refinement for opinion mining) System Phase:

Pre-processor:

This phase conducts morphological analysis of the review data including POS tagging, splits a compound sentence into multiple sentences, and performs stop word removal and stemming.

1) Morphological Analysis: In the initial step of pre-processing, FEROM eliminates the unnecessary content, such as tags, dates, and reviewer names, from the collected review data. Then, to extract noun phrases from the review data as feature candidates, NLP is used to perform morphological analysis, including POS tagging [7]. In general, morphological analysis is an essential component of natural language processing, dealing with the componential nature of words which are composed of morphemes. Morphological analysis recognizes the words that the text is made up of and identifies their part of POSs.

2) Sentence Splitting: - Sentence splitting is a process for segmenting a compound sentence containing conjunctions into several simple sentences. Sentence splitting is necessary because compound sentences may contain several features, each of which may represent different opinion information. Our method of splitting a compound sentence is carried out by recognizing a complete clause which is comprised of a noun phrase and a verb phrase. When several complete clauses exist in a sentence, the connective words ('and', 'or', 'but,' and so on) and the comma (',') are used to segregate them. Hence, the first step of sentence splitting is to divide the input sentence into several candidate complete clauses by simply separating the sentence when a conjunctive word or the comma is encountered. The next step is to examine each candidate clause to see if it is complete, namely, containing both a noun phrase and a verb phrase. A candidate clause that meets this condition is recognized as a complete clause. On the other hand, a candidate clause that does not satisfy this requirement is not complete, and hence is regarded as a component of the previous complete clause.

Feature Extractor Phase:

The feature extractor extracts product features from pre-processed review data. Feature extraction proceeds in three phases: feature selection selects a candidate feature in a sentence by looking for a noun phrase, opinion information extraction finds an opinion phrase that is associated with the candidate feature, and opinion phrase conversion replaces an opinion phrase expressed using a negative term with its antonym.

1) **Feature Selection: -** After sentence splitting, we can assume that each sentence contains opinion information about a single feature. In general, a feature in a sentence is in the form of a noun phrase, so feature selection normally proceeds by selecting noun phrases

2) **Opinion Information: -** Opinion information is expressed through opinion phrase. The opinion information in a sentence is expressed with negative terms such as 'not,' 'no,' and 'hardly.' In this case, the orientation of the opinion about the feature is the opposite of the meaning of the corresponding opinion phrase. Hence, for correct analysis, FEROM employs the opinion phrase conversion process that replaces an opinion phrase expressed using a negative adjective phrase with its antonym using Word Net [8]. For example, the sentence "the picture quality is not good" is changed into "the picture quality is bad."

3) **Opinion Phrase Conversion: -** The algorithm for opinion phrase conversion first determines whether a negative term exists in a sentence, and then calculates the distance between the opinion phrase and the negative term by counting the number of

words between them. If the distance is smaller than the threshold α, the opinion phrase is replaced by its antonym.

4) Feature Refiner: The feature refiner reduces the number of features by merging candidate features with the same or similar meanings, defined as homogeneous features. The feature refiner recognizes homogenous features by exploiting the feature ordering process that synchronizes the word orders of the features to detect synonymous feature candidates and the feature containment checking process that examines the subset superset relationship between the features to check for similarity between them.

1) Homogenous Feature Recognition: - Features are defined as features with the same or similar meanings. FEROM determines whether the extracted features are homogeneous using the synonym relation process in WordNet. In other words, if there is a synonym similar and homogeneous, as shown in Figure 5. relation between two features, they are regarded as

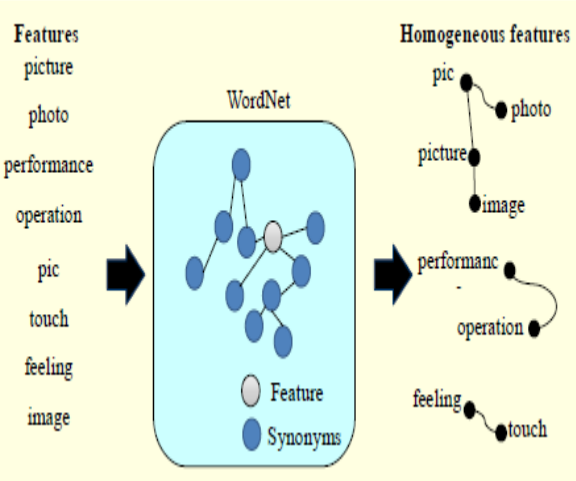

Figure 5. Detecting homogeneous features using synonym relation in Word Net.

However, in the case of a compound noun, the process of recognizing homogeneous features is less obvious. In linguistics, two compound nouns that are comprised of the same nouns but with different word order might possess different meanings. Hence, to determine the homogeneity of 'compound noun'-based features, we consider three ways to determine whether or not the features consisting of *n*-words represent the same meaning. FEROM describes all the above features, that is, 'picture quality,' 'quality picture,' 'photo quality,' and 'quality photo,' as homogeneous. This implies that, for compound nouns, FEROM emphasizes the synonymy of individual words and ignores the word orders [9].

2) Feature Merging: - Finally, the feature merging process merges homogeneous features into a representative feature and also prunes the feature candidates that have significantly low frequencies and very small amounts of related opinion information. Once several features are determined to be homogeneous, the next step is to merge them into a single feature.

Previous research on feature-based opinion mining has not had good results due to drawbacks, such as selecting a feature considering only syntactical grammar information or treating features with similar meanings as different. To solve these problems, this paper proposes an enhanced feature extraction and refinement method called FEROM that effectively extracts correct features from review data by exploiting both grammatical properties and semantic characteristics of feature words and refines the features by recognizing and merging similar ones [10].

The output of the FEROM system is given to Feature and opinion identification module we represents the dependency relations between a pair if words w1 and w2 is as relation type(w1,w2), in which w1 is called head or governor and w2 is called dependent or modifier. This may be direct or indirect Relation type id. In direct, one word depends on other directly and in indirect on through other word or both of them depends on third word indirectly. As information component is defined as < f, m, and o >. This module represents rule based system for formal text as in paper [6]. For informal text for example we are in a dependency relation R, if there exists an abbrev (w1, w2) relation such that POS (w1) = NN, POS (w2) = JJ, and w1 and w2 are not stop-words then w2 is assumed to be an opinion and w1 as a feature. After identifying feature and opinion final opinion is given to user.

V CONCLUSIONS

In opinion mining, feature extraction is important since the customers do not normally express their product opinions holistically but separately according to its individual features. However, previous research on feature-based opinion mining has not had good results due to drawbacks, such as selecting a feature considering only syntactical grammar information or treating features with similar meanings as different. As well as existing methods of feature-based opinion mining not work for informal text document. To solve these problems, our proposed system called enhanced feature extraction and refinement for Informal Text Mining (FITM) can be used for sentiment analysis classification. That effectively extracts correct features from review data by exploiting both grammatical properties and semantic characteristics of feature words and refines the features by recognizing and merging similar ones. FITM is highly effective at extracting and refining features for analysing customer review data and eventually contributes to accurate Informal Text Mining. FITM is a proper method for opinion mining by employing an enhanced scheme of feature extraction and refinement to analyse customer review data. By using this method we not only overcome the disadvantages of existing feature-based opinion mining but also achieve the Informal Text Mining.

REFERENCES

[1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, 2008, pp. 1-135.

[2] A Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews", Proceeding, National Conference on Human Language Technology Empirical Methods Natural Language Processing, 2005, pp. 339-346.

[3] Samaneh Moghaddam & Martin Ester, "Opinion Mining in Online Reviews: Recent Trends" Simon Fraser University Tutorial at WWW2013, May 14th 2013, 22nd International World Wide Web Conference.

[4] Ahmad Kamal, Muhammad Abulaish , Tarique Anwar, "Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Source", WIMS '12, June 13-15, 2012 Craiova, Romania. Copyright 2012 ACM 978-1-4503-0915-8/12.

[5] Fadi Abu Sheikha and Diana Inkpen, "Learning to Classify Documents According to Formal and Informal Style", LiLT Submitted, March 2012, Published by CSLI Publications.

[6] Marie-Catherine de Marneffe and D.Manning, Stanford typed dependencies manual. Revised for Stanford Parser v.1.6.9 in September.

[7] NLProcessor–Text Analysis Toolkit, 2000. http://www. Infogistics.com/textanalysis.

[8] G. Miller et al., "Introduction to WordNet: An On-line Lexical Database," Int. J. Lexicography, vol. 3, no. 4, 1990, pp. 235-244.

[9] Hana Jeong et al.," FEROM: Feature Extraction and Refinement for Opinion Mining" ETRI Journal, Volume 33, Number 5, October 2011.

[10] B. Liu and M. Hu, "Mining Opinion Features in Customer Reviews," Proc. 19th Nat. Conf. Artificial Int., 2004,pp.755-760.

