

# EXTRACTING BUSINESS INTELLIGENCE FROM ONLINE PRODUCT REVIEWS

<sup>1</sup>Soundarya.V, <sup>2</sup>Siddareddy Sowmya Rupa, <sup>3</sup>Sristi Khanna, <sup>4</sup>G.Swathi,  
<sup>5</sup>Dr.D.Manjula

<sup>1,2,3,4,5</sup> Department of Computer Science And Engineering, Anna University, Chennai,  
India

Email: <sup>1</sup>soundar\_riya@yahoo.co.in, <sup>2</sup> mail2rupas@gmail.com,  
<sup>3</sup>khanna.sristi@gmail.com, <sup>4</sup>swate.reddy@gmail.com, <sup>5</sup>manju@annauniv.edu

## ABSTRACT

*The project proposes to build a system which is capable of extracting business intelligence for a manufacturer, from online product reviews. For a particular product, it extracts a list of the discussed features and their associated sentiment scores. Online products reviews and review characteristics are extracted from www.Amazon.com. A two level filtering approach is adapted to choose a set of reviews that are perceived to be useful by customers. The filtering process is based on the concept that the reviewer generated textual content and other characteristics of the review, influence peer customers in making purchasing choices. The filtered reviews are then processed to obtain a relative sentiment score associated with each feature of the product that has been discussed in these reviews. Based on these scores, the customer's impression of each feature of the product can be judged and used for the manufacturers benefit.*

## 1. INTRODUCTION

With the rapid growth of the Internet, product related word-of-mouth conversations have migrated to online markets, creating active electronic communities that provide a wealth of information. Reviewers contribute time and energy to generate reviews, enabling a social structure that provides benefits both for the users and the firms that host electronic markets. In such a context, "who" says "what" and "how" they say it, matters. On the flip side, a large number of reviews for a single product may also make it harder for individuals to track the gist of user's discussions and evaluate the true underlying quality of a product. Recent work has shown that the distribution of an overwhelming majority of reviews posted in online markets is bi modal. Reviews are either allotted an extremely high rating or an extremely low rating. In such situations, the average numerical star rating assigned to a product may not convey a lot of information to a prospective buyer or to the manufacturer who tries to understand what aspects of its product are important. Instead, the reader has to read the actual reviews to examine which of the positive and which of the negative attributes of a product are of interest.

## 2. LITERATURE SURVEY

The task of assessing product reviews is related to these broader areas of research: The FAST-feature selection algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features [1]. A count-prediction based algorithm is proposed, which estimates the counts of item sets by predictive models to find frequent item sets out. The predictive models are constructed based on the data in the data stream and serve as a description of the concept of the stream [2]. To solve the existing problem of network text formalization in subjective and objective text classification, a machine learning classification method based on network informal language (NIL) is proposed. Firstly, a network informal dictionary is constructed by writing a web crawler to collect informal words which can be divided into two categories: typical type and fuzzy type. Then, different methods are put forward to formalize the informal network text based on the two types of informal words. Finally, we adopt the Native Bayes classifier and Sequential Minimal Optimization classifier to distinguish subjectivity and objectivity of the text [3]. Another model includes the first step which is to decide the polarity of a review by performing sentiment analysis. We can then classify the review based on the polarity [4]. In another method taking the characteristic of the product reviews into account, this method firstly extracts the candidate keywords, and then filters out noise keywords based on the rules. And then extend these keywords based on the correlative words recognition. This method finally realizes the hierarchical product review detection method based on these keywords. The experimental results show that the method proposed in this paper is successful [5]. In another method, the equivalence redundancies of fuzzy items and related theorems as a new concept for fuzzy data mining are defined. Then, a basic algorithm is proposed based on the Apriori algorithm for rule extraction utilizing the equivalence redundancy of the fuzzy items based on redundancy concepts of fuzzy association rules [6].

## 3. ARCHITECTURE

The overall system design is depicted in *Figure-1*. There are 3 main models: Predictive, Explanatory and Feature Model. Predictive and Explanatory model together form the filtering model. The explanatory module has further sub modules. The input to the system is the amazon web pages which is downloaded and extracted using Perl to form the required data set. Subsequently the product reviews and review characteristics obtained are the input to the first module. Finally, the output is the business intelligence that has been inferred from the processing of this data.

The subsequent sections explain each module in detail. **Section 4** explains the Predictive Model, **Section 5** describes the Explanatory Model in detail, **Section 6** explains the Feature Model, **Section 7** discusses the conclusion and future work.

## 4. PREDICTIVE MODEL

In the first stage of the project, a decisive predictive model is built. This classifies an unseen review as helpful or not. This constitutes to the first level of filtering in the two-level review filtering approach.

The helpfulness of each review in our data set is defined by the votes given by the peer customers, they decide whether a review is helpful or not. Prior research has demonstrated an association between the number of helpful votes of reviews and the subsequent sales of the product on the website, to the extent that better reviews receive more number of total votes and subsequently more number of helpful votes. Research has also demonstrated that as reviews are read and voted by more number of people, it signifies that the review reveals the true underlying quality of the product. Since the number of online reviews for each product is huge in number and it is impractical and an unnecessary wastage of resources to process each review, this approach of filtering reviews based on just votes is practical. This model has been implemented using decision trees, where expected helpfulness values of competing alternatives are calculated.

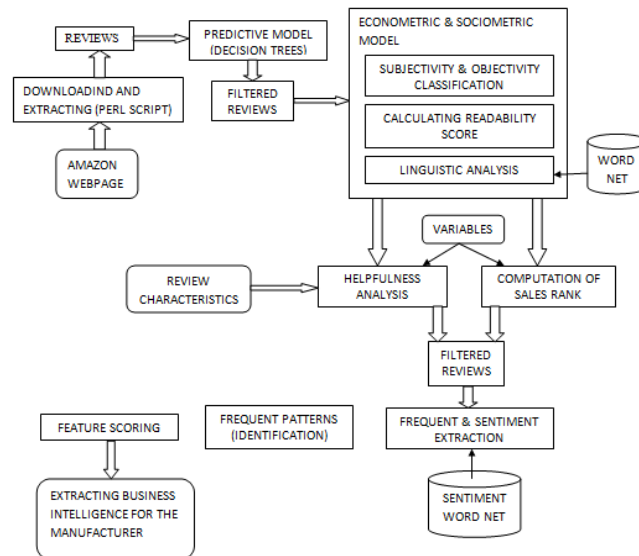


Figure 1.

**Decision Trees** - Decision trees have been selected for this model because they are simple to handle and interpret. They can handle both numerical and categorical data, are robust and make it possible to validate a model using statistical tests. They are chosen mainly because they perform well with large data in a short time.

Since customers tend to read reviews with more number of votes and a higher helpfulness ratio, these two have been chosen as variables for the decision tree. The helpfulness ratio is the number of helpful votes divided by the total number of votes.

After careful analysis of several product reviews, the threshold for the both the variables were chosen. The root of the tree, branches into two, the total votes being greater than 50 and lesser than 50. If the total votes are greater than 50 and the helpfulness value is greater than 0.7, then the review is chosen. If the total votes are in the range 10 – 50, and the helpfulness ratio is greater than 0.5 then also the review is selected. But if the total votes are less than 10 and only if the helpfulness ratio is equal to 1, the review is assured to be filtered as a helpful review. The tree has been constructed as shown below:

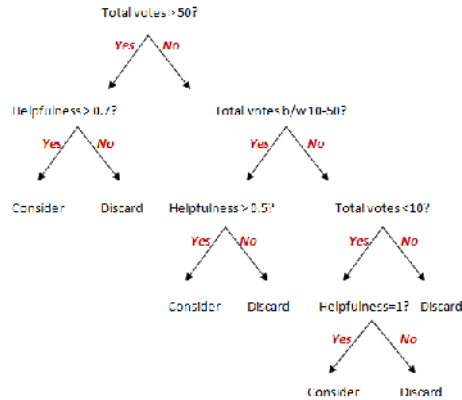


Figure-2

The output of the predictive model is a set of reviews that are perceived to be useful by the product customers. The rest of the reviews are discarded and only the filtered reviews are provided as an input to the next filtering step.

## 5. EXPLANATORY MODEL

The helpfulness of the filtered reviews obtained from the predictive model is determined by means of the explanatory model using metrics like subjectivity probability, normalized spelling error count, readability score, disclosures and rating. In this model we focus mainly on the textual aspects of the review to see how they affect the helpfulness of the review [1].

### 5.1 READABILITY SCORE CALCULATION

Easy-reading text improves comprehension, retention, and reading speed. A review that can be easily read is expected to influence a large number of customers. The filtered reviews are considered for the calculation of the readability score. There are numerous metrics for measuring the readability of a text, and while none of them is perfect, the computed measures correlate well with the actual difficulty of reading a text. To avoid idiosyncratic errors peculiar to a specific readability metric, a set of metrics are computed for each review and then averaged. The length of each review is measured; the number of sentences, words, characters and syllables are counted. These are used for the calculation of various metrics [2]. The readability metrics used are:

- Automated Readability Index

- Coleman-Liau Index
- Flesch Reading Ease
- Flesch-Kincaid Grade Level
- Gunning fog index
- SMOG

The readability score is calculated by taking the average of these indices.

## **5.2 LINGUISTIC ANALYSIS**

A review that is spelling error free is easier to read and is more impactful than a review that has spelling errors. To measure the spelling errors, a spell checker is used. The number of the spelling errors in each review is calculated. This is done by ignoring: capitalized words, words with numbers in them and the top hundred most frequent non-English words that appear in reviews, such as brand names and terminology words. Once the spelling error count is calculated for a review, it is normalized by dividing it by the length of the review (in characters).

## **5.3 SUBJECTIVITY AND OBJECTIVITY CLASSIFICATION**

Beyond spellings and readability, we also expect that there are stylistic choices that affect the perceived helpfulness of a review. There are two types of listed information, from the stylistic point of view. There are reviews that list "objective" information, listing the characteristics of the product, and giving an alternate product description that confirms (or rejects) the description given by the merchant. The other types of reviews are the reviews which are "subjective", containing sentimental information, in which the reviewers give a very personal description of the product, and information that typically does not appear in the official description of the product [3].

As the first step towards understanding the impact of the style on review helpfulness, existing literature on subjectivity estimation from computational linguistics is used. Pang and Lee described a technique that identifies which sentences in a text convey objective information and which of them contain subjective elements [4]. They applied their techniques on a movie review data set, in which they considered the movie plot as objective information, and the information that appeared in the reviews as subjective. In this scenario, the same paradigm is followed: the information that appears in the product description is considered as objective information and everything is taken as subjective.

## **5.4 REVIEW DISCLOSURE**

Social information about reviewers themselves is likely to be an important predictor of consumer's buying decisions. On many sites, social information about the reviewer is as prominent as product information. Given the extent and salience of available social information regarding product reviewers, it seems important to control for the impact of such information on online product sales and review helpfulness.

Amazon has a procedure by which reviewers can disclose personal information about themselves. There are several types of information that users can disclose: their real name, location,

nickname, and hobbies. This information is encoded as a binary value. An additional variable labeled “disclosures” is created; this variable captures each instance where the reviewer has engaged in one or more of the four kinds of self-disclosure. On the presence of a disclosure the variable is assigned a value of 0.5, otherwise the variable is assigned a value of zero.

## **5.5 REVIEW RATING**

For each review, the review rating of the product given by the reviewer is extracted. It is retrieved along with the other characteristics of the review. The rating that a reviewer allocates to the product is denoted by the number of stars, marked on a scale of 1 to 5. A single star rating signifies the lowest score while 5 stars are assigned to convey the highest rating. It is observed that the helpfulness is directly related to the star rating awarded to the review [6].

## **5.6 HELPFULNESS CALCULATION**

Taking an aggregate of the above mentioned calculated values, the helpfulness score for each review is computed. This value takes into account: the average readability score of the review, the product rating in the review, the presence or absence of a disclosure by the reviewer, the normalized spelling error count of the review and the subjectivity probability score of the review.

The final helpfulness value of each review is calculated using the formula:

$$\text{Helpfulness} = \text{ReadScore} + \text{Sprob} + \text{Disclosures} + \text{Rating} - \text{Ncount}$$

These values are sorted and the top 100 reviews are chosen to be filtered. These 100 reviews are the output of the overall filtering system and the input to the next model.

## **6. FEATURE MODEL**

The main goal of extracting the features and their related sentiment is to provide to the manufacturer the features that have a positive or negative impact in his sales. As a result he can concentrate on those features to improve his sales. To the manufacturers, these individual weaknesses and strengths are equally important to know, or even more valuable than the overall satisfaction level of customers.

This involves three steps namely

- Extracting topic specific features
- Extracting sentiment of each sentiment bearing phrase along with polarity
- Feature – Sentiment Association

### **6.1 POS TAGGING**

The first step of the feature model is to POS tag all the reviews that have been obtained after the two levels of filtering. The reviews are inputted to the Stanford POS tagger, which outputs the reviews along with the POS tag for each lexical item i.e. each word, punctuation mark and

number. Each lexical item is succeeded by its POS tag. It is in the format: *<lexical\_entry>*  
*<POS\_tag>*

## 6.2 FEATURE LIST CREATION

The aim of the Feature list creation sub module is to create a product specific feature list dynamically. Features present in the reviews are mapped on this list; all further analysis is done only on features present in this list.

A sample set of the POS tagged product reviews are taken as input for this sub module. A set of Phrase Patterns are identified. On occurrence of a pattern match, the matching phrase or word is taken as a possible feature. Once all the possible features have been extracted, the high frequency features are selected to form the feature list for that specific product.

Features of a topic satisfy one of the following relationships:

- A part-of relationship with the given topic.
- An attribute-of relationship with the given topic.
- An attribute-of relationship with a known feature of the given topic.

Based on the observation that feature terms are nouns, only noun phrases are extracted. The candidate feature terms are obtained from the following noun phrases:

### 6.2.1 BASE NOUN PHRASES (BNP):

The phrase patterns are: *NN*, *NN NN*, *JJ NN*, *NN NN NN*, *JJ NN NN*, and *JJ JJ NN* where *NN* and *JJ* are the POS tags for nouns and adjectives respectively. On occurrence of a longer BNP, the shorter BNPs are ignored to avoid repetition of features.

### 6.2.2 DEFINITE BASE NOUN PHRASES (DBNP):

dBNP further restricts candidate feature terms to BNPs that are preceded by the definite article “**the**”. Given that a document is focused on a certain topic, the definite noun phrases referring to topic features does not need any additional constructs such as attached prepositional phrases or relative clauses for the reader to establish its referent.

### 6.2.3 BEGINNING DEFINITE BASE NOUN PHRASES (BBNP):

bBNP refers to a dBNP occurring at the beginning of sentences followed by a verb phrase. This heuristic is based on the observation that, when the focus shifts from one feature to another, the new feature is often expressed using a definite noun phrase at the beginning of the next sentence.

In the candidate feature term list created above, the frequency of each candidate feature term is calculated. The features are then sorted based on their frequencies. The features with a higher frequency are chosen to form the feature list for that particular product.

### 6.3 SENTIMENT LEXICON

Sentiment about a subject is the orientation (or polarity) of the opinion on the subject that deviates from the neutral state. Sentiment that expresses a desirable state has positive (or '+') polarity, while one representing an un-desirable state has negative (or '-') polarity. The system uses sentiment terms defined in the sentiment lexicon.

The sentiment lexicon contains the sentiment definition of individual words in the following format:

*<lexical entry ><POS ><sent category ><sent score >*

where,

- *lexical entry*: the term that has sentimental connotation
- *POS*: the required POS tag of the lexical entry
- *sentiment category*: positive(+) or negative(-)
- *sentiment score*: the score given to the lexical entry, on a scale of 1 to 5

There are about 2500 sentiment term entries including about Nouns, adjectives, verbs, adverbs and propositions that have been included in the sentiment lexicon. The sentiment lexicon serves as a lookup database for each instance of sentiment words present in the product reviews.

### 6.4 FEATURE AND SENTIMENT SCORING

From each sentence of each review, the following phrases are identified:

#### **Adjective Phrase:**

*<Verb ><Adjective >, <Adverb ><Adjective >*

#### **Prepositional Phrase:**

*IN NN PRP, IN DT NN, IN IN PRP, IN IN DT NN, IN DT PRP NN, IN IN DT JJ NN*

#### **Subject Object Phrase:**

The BNPs of the sentiment words are Subject Object Phrases.

Initially, all the features are assigned a score of zero. Once all the phrases have been identified, the features and sentiments occurring in them are extracted. The features are matched to the feature list; the associated sentiment is matched to the sentiment lexicon to obtain the polarity and sentiment score from the database. The sentiment score obtained is added or subtracted to the current sentiment score of the associated feature. This process is repeated for all the identified phrases. Each Feature now has its own cumulative score [7].



*Multiple sentiment terms associated with a particular feature term:* The score associated with all the sentiments is acquired and the net total is calculated, this is added or subtracted to the existing feature score.

*Single sentiment term is used to describe multiple features:*

The sentiment score along with its polarity is added to each of the features individually.

*Feature term is not explicitly stated but the sentiment is expressed:*

The mapping of the sentiment to the feature term is done with the help of words like: **It, that, which** and **this**. The feature that has been previously discussed is taken as the current feature under consideration. The sentiment value is added to that feature.

#### **Sentiment words preceded by negative words:**

When sentiment words are preceded by negative words such as **not, no, never, hardly, seldom,** or **little**, then the polarity of the sentiment is reversed.

#### **Sarcastic sentences:**

The presence of sarcastic sentences reverses the polarity of the sentiment expressed. The presence of multiple occurrences of question marks, exclamation marks and capitalized words is taken as a test to identify if the phrase under consideration is sarcastic or not. Once a sarcastic sentence has been identified, the polarity of its calculated sentiment score is reversed [8].

#### **Sentiment words are sometime preceded by words like too and very:**

In such cases, the polarity of these special words depends on the polarity of the sentiment word succeeding them. Thus the sentiment score gets increased by 2 units either on the positive scale or the negative scale depending on the polarity of the sentiment word. When such cases are preceded by the negative words discussed above, first the net sentiment score is calculated and then the polarity is reversed.

Finally a list of features and its corresponding sentiment score is obtained. This list is sorted based on the sentiment score. The highest scoring features are the ones that are the most liked features and have been spoken about positively. The lowest scoring features are the features that have been disliked and have been spoken about negatively. The feature terms with neither a positive nor a negative score i.e. features with a sentiment score equal to zero have three possibilities:

- The particular feature term has not been spoken about in the review.
- The feature term has been mentioned in the review, but neither a definite positive nor negative comment has been made.
- The feature term has a set of positive and negative comments but they cancel out each other leading to a total score of 0.

All these features are classified as neutral features and are of less importance to the manufacturer.

## 7. RESULTS AND PERFORMANCE

The performance analysis is carried out on a wide range of products. Three types of products are chosen: Mobile Phones, Tablets and Television Sets for the Samsung manufacturer. Further, for each type of product, five different models are chosen. For each model, a set of reviews are analysed. On obtaining a set of product reviews for the product taken under consideration, the system is run for that input. The same computation is carried out manually as well. The two values obtained from the system and manually are compared and analysed. The deviation is calculated for both Subjectivity Probability and Feature Score. From the actual and calculated feature score, the accuracy % is calculated for each product type. Using the deviation in the computed feature score the precision range is calculated, it is the range of the minimum deviation to the maximum deviation, calculated for each single feature of the product under consideration. All this is show in Table below.

Table 1.

	Mobile Phones	Tablets	Televisions
Sprob deviation	0.134	0.154	0.138
Accuracy %	93.47%	94.06%	92.67%
Precision Range	0-4	0-3	0-3

## 8. CONCLUSIONS AND FUTURE WORK

**Conclusion** - The project deals with two main ideas – to filter reviews according to their perceived usefulness and extract business intelligence from them. The filtering is done in two stages. First, the unseen reviews are chosen based on their helpfulness ratio obtained from peer customer votes. Next, the reviews are processed to get a cumulative score based on readability, spelling errors, review rating, subjectivity probably and disclosure of the reviewer. The highest scoring reviews are filtered. This approach ensures that the chosen reviews have useful content. The second module deals with mining the reviews for opinions. The novelty here is that instead of finding the overall sentiment of each review we are finding a sentiment score associated with each feature of the product. The project works for any product by dynamically creating the feature term list from the given input. The system is comprehensive as it includes taking customer reviews, analyzing them at multiple levels and then extracting the business intelligence. Special cases like implicit negation have also been handled.

**Future Work** - Words falling under the category of word sense disambiguation do not get included in the spelling error count. The spelling of the word is correct but the word is used out of context. This case can be handled. In case a review has sentiments associated with another product, they also get mapped on the features of the current product being discussed. This issue can be rectified.

## ACKNOWLEDGEMENT

We are glad to take this opportunity to cordially acknowledge a number of people who provide us with a great support, First we would like thank my guide Dr.D.Manjula, who give full support to do and guide me a lot to do my work, I like to thank my management for their constant support in doing my research work.

Furthermore, we would like to thank all our family members, Lectures, for their timely help in doing my work, by their generous support and encouragement throughout the life.

## REFERENCES

- [1] Qinqin Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 25, no. 1, January 2013.
- [2] Chao-Wei Li and Kuen Fang-Jea, "Using Count Prediction Techniques for Mining Frequent Patterns in Transactional Data Streams," *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, 2012.
- [3] Chupin Chao, Wenbao Jiang, "Study on the Subjective and Objective Text Classification and Pretreatment of Chinese Network Text," *2012 4th International Conference on Intelligent Human Machine Systems and Cybernetics*, 2012.
- [4] Arun Manicka Raja M., Godfrey Winster S., Swamynathan S, "Review Analyzer: Analyzing Consumer Product Reviews from Review Collections," *IEEE*, 2012.
- [5] Zhao Hua, Zeng Qingtian, Sun Bingjie, Ni Weijian, "Hierarchical Product Review Detection Based on Keyword Extraction," *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, 2012.
- [6] Toshihiko WATANABE, Ryosuke Fujioka, "Fuzzy Association Rules Mining Algorithm Based on Equivalence Redundancy of Items," *2012 IEEE International Conference on Systems, Man, and Cybernetics*, 14-17 October 2012, COEX, Seoul, Korea.
- [7] Anindhya Ghose and Panagiotis G. Ipeirotis, Member IEEE, "Estimating the Helpfulness and Economic Impact of product reviews: Mining text and reviewer characteristics," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 10, October 2011.
- [8] N. Hu, P. A. Pavlou, and J. Zhang, "Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication," *Proc. Seventh ACM Conf. Electronic Commerce (EC 06)*, pp. 324–330, 2006.
- [9] A. Ghose and P. G. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting the Usefulness and Impact of Reviews," *Proc. Ninth Intl Conf. Electronic Commerce (ICEC 07)*, pp. 303–310, 2007.
- [10] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically Assessing Review Helpfulness," *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP 06)*, pp. 423–430, 2006.
- [11] S.-M. Kim and E. Hovy, "Determining the Sentiment of Opinions" *20th Intl Conf. Computational Linguistics (COLING 04)*, pp. 1367–1373, 2004.
- [12] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. 12th Intl Conf. World Wide Web (WWW 03)*, pp. 519–528, 2003.
- [13] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Ann. Meeting of the Assoc. for Computational Linguistics (ACL 04)*, pp. 271–278, 2004.
- [14] Dmitry Davidov, Oren Tsur and Ari Rappoport, "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon," *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp 107-116, Uppsala, Sweden, 15-16 July 2010.

## **BIOGRAPHY**

V.Soundarya, received the Bachelor Degree of Engineering in Anna University, Chennai and Master Degree of Engineering in Anna University of Technology in Trichy, and doing her Research as part time scholar in Anna University, Chennai and working as Associate Professor in Dhanalakshmi Srinivasan College of Engineering and Technology, Chennai. She had published papers in conference and journals and attended some national conferences and workshops.

Siddareddy Sowmya Rupa, Sristi Khanna, G.Swathi are just completed their Bachelor of Engineering in Anna University, Chennai. They had published one paper in journal and attended two conferences.

Dr.D.Manjula received the PhD degree in computer science from the Anna University of Chennai in 2004. Currently, she is working as an associate professor in the Department of Computer Science and Engineering at Anna University, Chennai. She has published more than 80 papers in refereed journals and conference proceedings. Her major research area includes data mining. She has worked on various data mining problems including classification, clustering, indirect association mining, transitional pattern mining, diverging pattern mining, review mining, data stream mining, and bioinformatics. She is a member of the IEEE.