# Feature Subsumption for Sentiment Classification of Dynamic Data in Social Networks using SCDDF

Jayanag. B[1], Vineela. K[2], Dr. Vasavi. S[3]
Department of Computer Science and Engineering, V. R. Siddhartha Engineering College
Vijayawada, India.

***Abstract*- The analysis of opinions till now is done mostly on static data rather than on the dynamic data. Opinions may vary in time. Earlier methods concentrated on opinions expressed in an individual site. But on a given concept opinions may vary from site to site. Also the past works did not consider the opinions at aggregate level.**

**This paper proposes a novel method for Sentiment Classification that uses Dynamic Data Features (SCDDF). Experiments were conducted on various product reviews collected from different sites using QTP. Opinions were aggregated using Bayesian networks and Natural Language Processing techniques. Bulk amount of dynamic data is considered rather than the static one. Our method takes as input a collection of comments from the social networks and outputs ranks to the comments within each site and finally classifies all comments irrespective of the site it belongs to. Thus the user is presented with overall evaluation of the product and its features.**

***Keywords- Sentiment classification, Natural language processing (NLP); opinions; features; Quick Test Professional (QTP); feature identification; sentiment prediction; summary generation.***

## I. INTRODUCTION

The present opinion mining is done statically only for a small set of data and the dependencies in the opinions are not considered for summarization. An architecture that could automatically process the comments, generate a generalized result out of the list of comments posted about a product by considering the dependencies could be useful to give a brief synopsis of the product. This becomes a real-life application, a completely automated solution that extracts the comments posted in a social network and categorizes them based on most prominent ranks. Thus it helps the user to know about the pros and cons of a product and its features based on the existing user's feedback with little effort.

The proposed architecture is based on opinion mining, a sub discipline within data mining and computational linguistics, refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content.

Many numbers of sites provide different comments on the products but viewing all of them become rather difficult so we evaluate them based on the ranks and present a generalized result. The opinions posted by various users in social networks are extracted, the comments are evaluated by dividing them into tokens and using the natural language processing techniques like POS (Parts Of Speech) tagging. Meanings are analyzed by using the web dictionary WordNet [7, 21]. The dependencies in the opinions are analyzed using the Bayesian Networks and the sentiment is predicted for those corresponding words. And finally based on the predicted word counts ranks are given to these sentiments and are summarized. System gives the cumulative rank and displays the string corresponding to it.

Section II presents related work on the present study. Section III presents our proposed system. Experimental results are given in section IV. Conclusion and Future Work are given in Section V and VI.

## II. RELATED WORK

Previous works concentrated on opinions in individual sites and also limited the data set to a single line comment or static data or a limit on number of characters. Those current studies are mainly focused on mining opinions in reviews and/or classify reviews as to only positive or negative based on the sentiments of the reviewers but not on relative degree of positive or negativeness. Detailed study on previous works can be found in [13].

Abbasi et al. [1] considered web forms, blogs and articles and used WordNet score but haven't considered the word dependencies. Ahmed Abbasi [2], worked on feature selection methods and considered Intelligent Feature Selection (IFS) approach that uses syntactic and semantic information to refine larger input features, but these formation modules need to be expounded on, and real-world knowledge bases could be considered.

Cardie et al. [3], concentrated opinion-oriented information extraction. They created opinion-oriented "scenario templates" for summary representations of the opinions expressed in a document, or a set of documents to perform question answering. They did not identify product features and user opinions on these features to automatically produce a summary.

Dave et al. [4], worked on semantic classification of reviews as positive or negative ones using the available corpus from web sites, where each review already had a class e.g., binary ratings or thumbs-up and thumbs-downs. Sentiment classifiers are build around them. However, the performance was limited because a sentence contains much less information than a review.

Gary Beverungen et al. [8], considered twitter posts and summarized them using clustering. Here the data set is limited as the twitter posts considered are not more than 140

characters. Hsinchu Chen et al. [9], considered only Wal-Mart data set statically and categorized the data as direct and indirect opinions.

Minqing Hu et al. [10], considered opinions posted by customers, identified the features and gave the sentiment without considering the dependencies in the opinions. Morinaga et al. [11], compared reviews of different products of one category to find about the target product. However, it does not summarize reviews, and it does not mine product features on which the reviewers have expressed their opinions.

B. Liu et al. [12] handbook categorized the Information into two types: facts and opinions. The features are classified as explicit features and implicit features. But the dependencies are not considered here.

In [15] research work, they improved the performance of calculations and classifications using linguistic rules and constraints. Here supervised and unsupervised learning techniques are used. Feature selection methods, Information Gain (IG) and Mutual Information (MI), were applied and compared. They have compared their work with Ding et al. [6] but the results shows that there is a fall in precision and recall rates which clearly state that these methods are not that accurate.

In [18], it takes one comment at a time, the dependencies in the text are not considered and also techniques used for sentiment classification are not mentioned

In [20], NLTK 2.0.1rc1 powered text classification process is done. When the text is entered it expresses whether the text is positive negative or neutral sentiment. It takes one comment at a time, but here the results are not so accurate.

Thus the existing works are limited to a particular site or a static data set. And the opinions are just classified as positive opinions and negative opinions without considering the dependencies. Naïve Bayes classifier is used for sentiment classification.

But the dependencies that exist within words used in the comments are not considered. Section III presents our proposed system for sentiment classification.

## III. PROPOSED SYSTEM

The proposed system is a unique system which takes the data dynamically, classifies, ranks are given. These ranks may vary with in time and comments posted. Comments considered here are about mobile phones, cameras and laptops. Using this system the user can know the pro's and con's about a product.

Figure 1 presents the SCDDF architecture of our proposed system. The full length description of proposed system can be found in [13].

### A. Preprocessing

Firstly comments are collected dynamically from the sites using web crawler [14] QTP. Then the data set collected is tokenized [17]. Stop words like "a", "this", "is", etc are removed and dependency words like "not", "no" are considered.

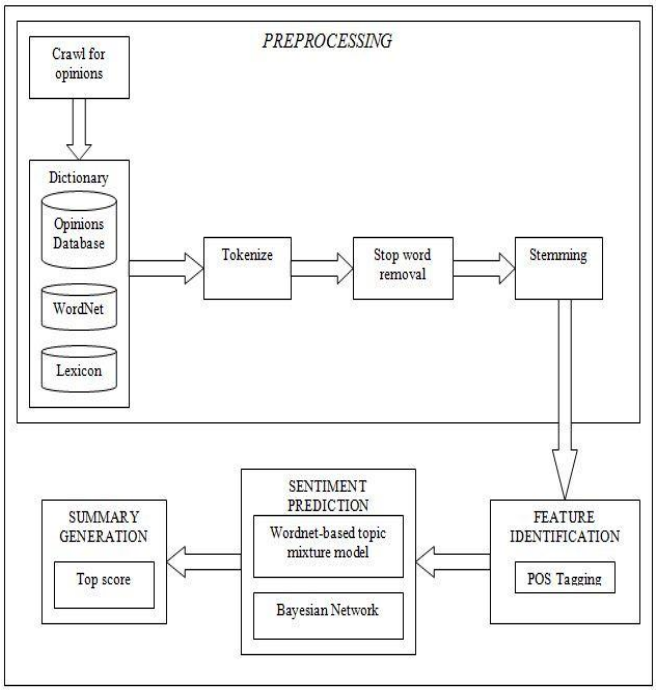


Fig. 1 Sentiment Classification for Dynamic Data Features (SCDDF).

At the end of preprocessing stemming is done. Stemming is the process where the words suffixes are removed. Porter stemmer [19] is applied for stemming. It is a 6 steps algorithm, where in each step the words are trimmed and the size of the data set will be reduced in each step.

### B. Feature Identification:

In this step features are identified for the comments collected. Features like "battery", "touch" etc are identified in this step. For this process POS (Parts Of Speech) –tagging is adverbs are identified for feature identification.

Feature Identification step:

$P(O,T)=\P_i P(t_{i-1}->t_i)p(w_i|t_i)$;

where,

P(O): Opinions

P(t):Tags

P(O, T): Opinions with tags

P(w): probability of getting a word from word net

### C. Sentiment Prediction:

In this step the sentiments for the comments i.e. positive and negative comments are predicted using wordnet [7, 21] and dependencies are resolved using the Bayesian network.

Example dependency comment:

*This is not a great mobile.*

Here "not" is a strong dependency word. Most of the works were dependencies are not considered says that the comment is a positive one as there is a positive word in it. But in actual sense it is a negative comment.

$$P(C \mid A \text{ ∧ } B) = P(C \mid B) \quad (1)$$

So by applying Bayesian networks these dependencies are resolved.

*D. Summary Generation:*

At last considering the scores obtained from sentiment prediction level the results generated are shown using statistical summary report. Statistical summary report consists of comments, features extracted for each comment, positive or negative score assigned along with the positive and negative label and rank of the product. Thus the user can evaluate the odds and outs of the product.

## II. RESULTS

This section presents the experimented results of our proposed work SCDDF. To evaluate the performance of sentiment classification, we adopted four indexes that are generally used in text categorization:[5] Recall, Precision, F-measure and Accuracy. Performance is measured using the following metrics.

Experimental results shows that our method has produced an accuracy of 0.9 after preprocessing, 0.91 after feature identification and 0.918 after sentiment prediction for the product mobile. This shows that after each level the results are more refined to get accurate results.

Precision ( P ) = #Correct / #Guessed

Recall ( R ) = #Correct / #Relevant

Accuracy ( A ) = #Correct / # Total posts ; and

F-measure ( F ) = 2*Precision*Recall/ (Precision + Recall)

Table 1 presents the results of SCDDF when evaluated on sample data set.

| size | Product | Pre-processing | | | | Feature Identification | | | | Sentiment Prediction | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | A | F | P | R | A | F | P | R | A | F |
| 105 | Mobile | 0.91 | 0.92 | 0.905 | 0.915 | 0.914 | 0.93 | 0.91 | 0.922 | 0.921 | 0.935 | 0.918 | 0.928 |
| 60 | Camera | 0.917 | 0.922 | 0.91 | 0.92 | 0.924 | 0.931 | 0.92 | 0.927 | 0.931 | 0.94 | 0.932 | 0.935 |
| 60 | Laptop | 0.93 | 0.932 | 0.912 | 0.931 | 0.932 | 0.934 | 0.93 | 0.933 | 0.94 | 0.941 | 0.934 | 0.941 |

Table 1: The results of SCDDF when evaluated on sample data set.

Precision

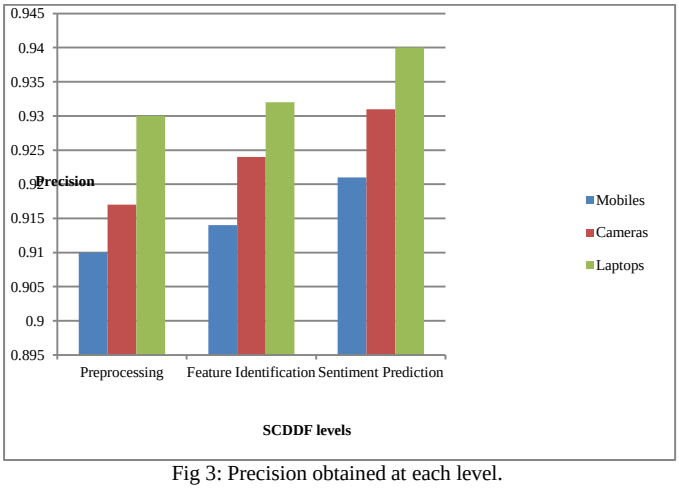


Fig 3: Precision obtained at each level.

Accuracy

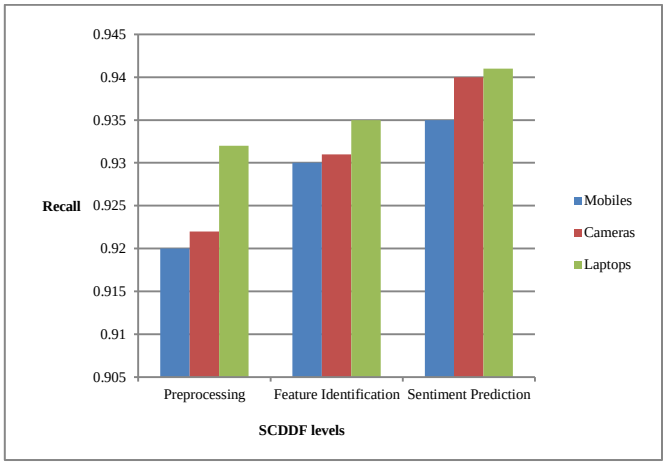


Fig 4: Recall obtained at each level.

Recall

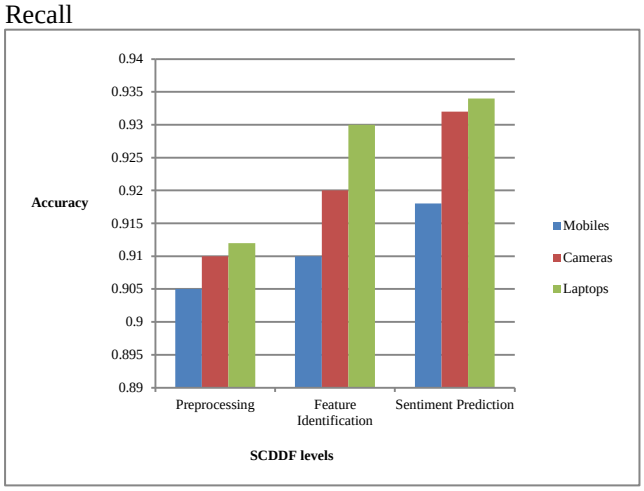


Fig 5: Accuracy obtained at each level.

F-measure

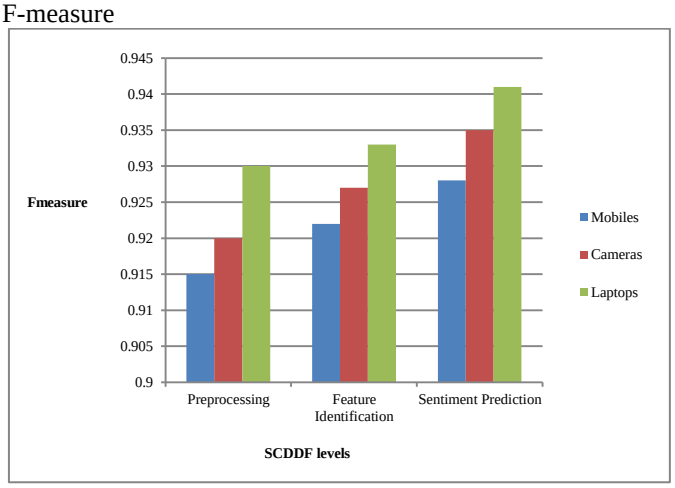


Fig 6: F-measure obtained at each level.

Similarly for camera and laptops data set, accuracy has been increased within each step. Figures 3, 4, 5,6 presents the comparison of the results with respect to each of the performance measure.

Table 2 presents the comparison of our method and online web tools. Figures 7,8,9 presents the snapshot of the execution of online tools.

| Sample input comments | Sentiment analyser [18] | Nltk [20] | SCDDF [13] |
|---|---|---|---|
| fall in love on this phone! elegant design & ideal specs. but i'm gonna buy the international version 'cos the brand logo is on top, on the bellow! :D | Overall sentiment is positive with probability of 0.985837 | The text is neutral. | pos 0.8901 |
| it really sucks that the T-Mobile Version is coming out in 6 days and will have the Ics straight out of the box and that's not fair. I think they should wait just like the rest of us at the back of the line and let the originals get the up-dates first. | Overall sentiment is negative with probability of 0.1854791 | The text is neg. | neg 0.3011 |
| Its not that good compared to iphone | Overall sentiment is positive with probability of 0.7626124 | The text is neg. | neg 0.2425 |
| Its good compared to iphone | Overall sentiment is positive with probability of 0.7626124 | The text is neg. | pos 0.7575 |

Table 2: Comparison of SCDDF with existing online tools.

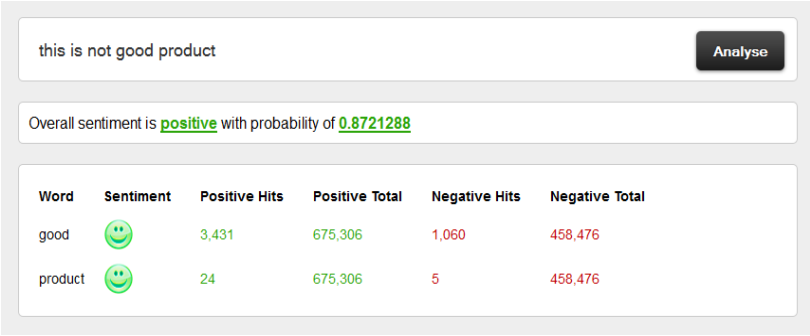


Fig 7: Snapshot of execution of [18]

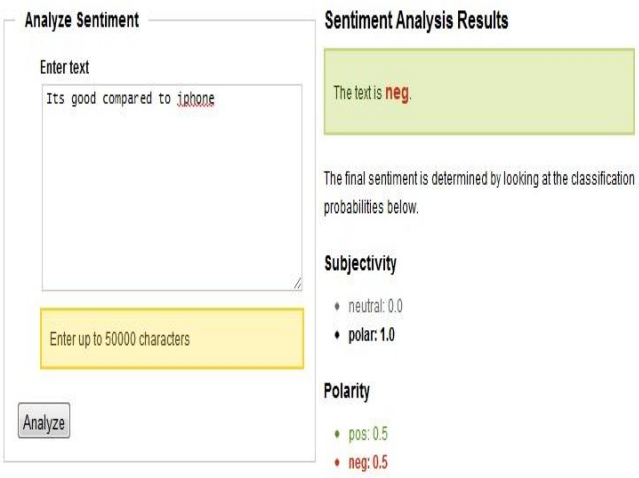


Fig 8: Snapshot of execution of [20]

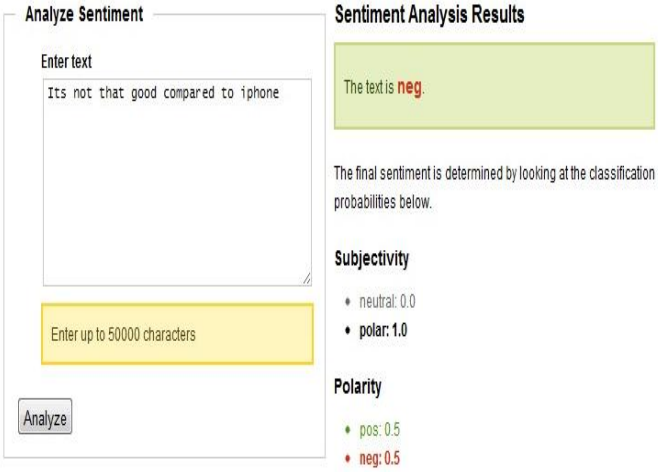


Fig 9: Snapshot of execution of [20]

The above results shows that our proposed method performs in the same way as existing tools in some cases where as performs better in other cases. The results clearly show that the existing tools neglected the dependencies within the comments. And also one comment at a time are analyzed. But our method takes dynamic data considering the dependencies using the Bayesian networks.

Table 3 Compares our work with existing works [6, 15].

Table 3 shows that there is a clear fall in Yanyan Meng values compared with Ding et al. methods. Our method SCDDF has increase in values when compared with [6,15]. In [15] they just identified the product features using techniques like document vector, sentence vector, intensification and sentence relation. These methods are useful to find the polarities and features. Ding et al. [6] applied the rule-based sentiment analysis technique which just says about the opinion orientations and product features. Both [6,15] neglected the dependencies in the words.

## IV. CONCLUSION

Feature subsumption for sentiment classification in social networks using natural language processing solves the problems in opinion mining and provides a novel approach for sentiment classification. It is a novel community-based evaluation that successfully captures the peculiarities of social networks.

However, the success of such an initiative eventually depends on the cooperation of the companies and institutions owning social network data, and on the agreement of enough organizations to participate in such a project.

## V. FUTURE WORK

Our future work concentrates on classifying the sentiments of messages posted within the social networks such as Facebook, Twitter. This is required as in the recent times government is planning to involve a 3rd person to analyze the comments posted over such networks. Even though this is violation to human right but protects the society without leading to unwanted situations. But basic human rights should not be destructed; in this situation without causing harm to anyone our method can find sensitivity of the messages posted in the network.

## REFERENCES

[1] Abbasi et al., "Affect Analysis of Web Forums and Blogs using Correlation Ensembles,"IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, 2008, pp. 1168–1180.

[2] Ahmed Abbasi, "Intelligent Feature Selection for Opinion Classification", University of Wisconsin-Milwaukee, - IEEE 2010.

[3] Cardie, C., Wiebe, J., Wilson, T. and Litman, D. 2003. Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. AAAI Spring Symposium on New Directions in Question Answering. 2003

[4] Dave, K., Lawrence, S., and Pennock, D., 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW'03.

[5] David D. Lewis. Evaluating and optimizing autonomous text classification systems. In SIGIR '95, pages 246--254, New York, NY, USA, 1995. ACM Press.

[6] Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining, In Proceedings of the Conference on Web Search and Web Data Mining (WSDM).

[7] Fellbaum, C. WordNet: an Electronic Lexical Database, MIT Press 1998.

[8] Gary Beverungen and Jugal Kalita, "Evaluating methods for summarizing Twitter Posts", WSDM'11, February 9-12,2011.

[9] Hsinchu Chen and David Zimbra, "AI and Opinion Mining", University of Arizona, - IEEE 2010

[10] Minqing Hu and Bing Liu, "Mining Opinion Features in Customer Reviews", American Association for Artificial Intelligence, 2004.

[11] Morinaga, S., Ya Yamanishi, K., Tateishi, K, and Fukushima, T. 2002. Mining Product Reputations on the Web. KDD'02.

[12] B. Liu, "Sentiment Analysis and Subjectivity," Handbook of Natural Language Processing, 2nd ed., N. Indurkhya and F.J. Damerau, eds., Chapman & Hall, 2010, pp. 627–666.

[13] B. Jayanag et al., "A Study on Feature Subsumption for sentiment classification in Social Networks using Natural Language Processing Techniques", Communicated to IJCA.

[14] Jeff Heaton, Programming Spiders, Bots, and Aggregators in Java, Publisher: Sybex, February 2002, ISBN: 0782140408

[15] Yanyan Meng Sentiment analysis: A study on product features, University of Nebraska.
[16] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[17] http://nltk.googlecode.com/svn/trunk/doc/api/nltk.tokenize-module.html
[18] http://sentiment.brandlisten.com/
[19] http://tartarus.org/martin/PorterStemmer/
[20] http://text-processing.com/demo/sentiment/
[21] http://wordnet.princeton.edu

| Data set [16] | Yanyan Meng methods [15] | | | | | | Ding et al. methods [6] | | | SCDDF [13] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | intensification | | | sentence relation | | | | | | | | |
| | P | R | F | P | R | F | P | R | F | P | R | F |
| Apex | 0.66 | 0.63 | 0.64 | 0.63 | 0.65 | 0.64 | 0.89 | 0.88 | 0.89 | 0.91 | 0.9 | 0.91 |
| CanG3 | 0.53 | 0.74 | 0.61 | 0.64 | 0.76 | 0.69 | 0.93 | 0.92 | 0.93 | 0.93 | 0.92 | 0.93 |
| Nikcool | 0.61 | 0.76 | 0.64 | 0.64 | 0.75 | 0.67 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| Nomp3 | 0.58 | 0.65 | 0.6 | 0.576 | 0.64 | 0.6 | 0.87 | 0.86 | 0.87 | 0.9 | 0.89 | 0.895 |
| No6610 | 0.66 | 0.79 | 0.72 | 0.68 | 0.82 | 0.74 | 0.95 | 0.95 | 0.95 | 0.952 | 0.95 | 0.95 |

Table : Comparison SCDDF with [6, 15]