ISSN No. 0976-5697

Volume 4, No. 10, September-October 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# Mining and Summarizing Opinion Features in Movie Reviews

J.Kanaka Priya*
M.Tech Scholar, Computer Science and Engineering
Gayatri Vidya Parishad College of Engineering(Autonomous)
Visakhapatnam ,Andhra Pradesh , INDIA.
kanakapriyajeyaraman@gmail.com

V.Adi Lakshmi
M.Tech Scholar, Computer Science and Engineering
Gayatri Vidya Parishad College of Engineering(Autonomous)
Visakhapatnam ,Andhra Pradesh , INDIA.
adilakshmi.sabella@gmail.com

G.V. Hindumathi
Asst Professor, Computer Science and Engineering
Gayatri Vidya Parishad College of Engineering(Autonomous)
Visakhapatnam ,Andhra Pradesh , INDIA.

***Abstract:*** Sentiment analysis, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as movies, products, services, organizations, individuals, issues, events, topics, and their attributes. [1]

For a particular movie, the number of reviews can be in hundreds or even thousands. This makes it difficult for a user to read and decide whether to watch the movie or not. So, we aim to mine and summarize all the user reviews of a movie. This summarization task is different from traditional text summarization because we only mine the features of the movie review on which the users have expressed their opinions and whether the opinions are positive or negative. We do not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Our task is performed in three steps: (1) mining features that have been commented on by users; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. This paper proposes several novel techniques to perform these tasks.

***Keywords:*** Text mining, sentiment classification, summarization, reviews.

## I. INTRODUCTION

With more and more common users becoming comfortable with the Internet, an increasing number of people are writing reviews. As a consequence, the number of reviews that a movie receives grows rapidly. Some popular movies can get hundreds of reviews at some large sites. This makes it very hard for a potential user to read them to help him or her to make a decision on whether to watch the movie.[6]

In this research, we propose to study the problem of *feature-based opinion summarization* of user reviews of movies. The task is performed in two steps:

a. Identify the features of the movie that users have expressed opinions on (called *opinion features*) and rank the features according to their frequencies that they appear in the reviews.
b. For each feature, we identify how many user reviews have positive or negative opinions. The specific reviews that express these opinions are attached to the feature. This facilitates browsing of the reviews by potential users.

Simple example to illustrate. Assume that we summarize the reviews of a particular movie,
***Movie_1***. Our summary looks like the following:
**Music:**
Positive: 253 <individual reviews>
Negative: 6 <individual reviews>
**Story**:
Positive: 134 <individual reviews>
Negative: 10 <individual reviews>
**…**

***Music*** **and** ***Story*** are ***opinion features***. There are 253 reviews that express positive opinions about the Music, and only 6 that express negative opinions. <Individual reviews> points to the specific reviews that give positive (or negative) comments about the feature. With such a feature-based opinion summary, a potential user can easily see how the existing users feel about the Movie. If he/she is very interested in a particular feature, he/she can drill down by following the <individual reviews> link to see why existing customers like it or what they complain about.

We are only interested in features of the movie that users have opinions on and also whether the opinions are positive or negative. We do not summarize the reviews by selecting or rewriting a subset of the original sentences from the reviews to capture their main points as in traditional text summarization.

## II. RELATED WORK

Our work is mainly related to two areas of research, text summarization and terminology identification. The majority of text summarization techniques fall in two categories: template instantiation and text extraction. Work in the former framework includes (DeJong 1982), (Tait 1983), and (Radev and McKeown 1998). They focus on the identification and extraction of certain core entities and facts in a document, which are packaged in a template. This framework requires background analysis to instantiate a template to a suitable level of detail. In [2], Morinaga *et al.* compare reviews of different products in one category to find the reputation of the target product. However, it does not summarize reviews, and it does not mine product

© 2010, IJARCS All Rights Reserved

features on which the reviewers have expressed their opinions. Although they do find some frequent phrases indicating reputations, these phrases may not be product features (e.g., "doesn't work", "benchmark result" and "no problem(s)"). In [3], Cardie *et al* discuss opinion-oriented information extraction. They aim to create summary representations of opinions to perform question answering. They propose to use opinion-oriented "scenario templates" to act as summary representations of the opinions expressed in a document, or a set of documents. Our task is different.

We aim to identify features and user opinions on these features to automatically produce a summary. Also, no template is used in our summary generation.

## III. PROPOSED WORK

Figure 1 gives the architectural overview of our opinion summarization system.

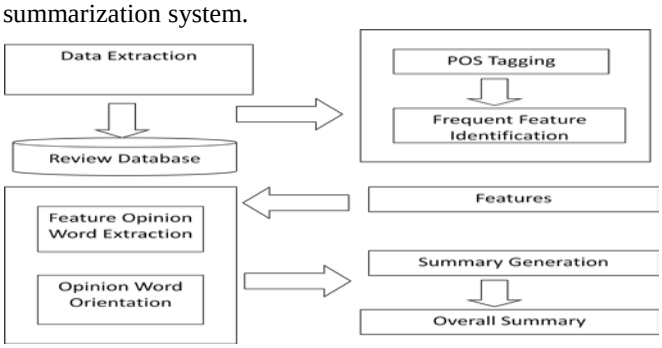


Figure 1: Feature-based opinion summarization

The inputs to the system are a movie name and an entry Web page for all the reviews of the movie. The output is the summary of the reviews. The system performs the summarization in three main steps : (1) mining features that have been commented on by users; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. These steps are performed in multiple sub-steps. Given the inputs, the system first downloads (or crawls) all the reviews, and put them in the review database. It then finds those "hot" (or frequent) features that many people have expressed their opinions on. After that, the opinion words are extracted using the resulting frequent features, and semantic orientations of the opinion words are identified with the help of WordNet [4]. Using the extracted opinion words, the system then finds those infrequent features. In the last two steps, the orientation of each opinion sentence is identified and a final summary is produced. Note that POS tagging is the part-of-speech tagging [5] from natural language processing, which helps us to find opinion features. Below, we discuss the algorithm.

## IV. PROPOSED ALGORITHM

***// Review Extraction***
***Step 1 :For every review r***
***Step 2:*** *Split r into sentences*
***Step 3: For every sentence s in r***
***Step 4 :*** *posTaggedSentence= PosTaggerInterfacer.Pos(sentence);*
***// Feature Extraction***
***Step 5 : For every word w in s***
***Step 6 : If word is a noun and not present in nouns hash then put w in hash table and set Count{word}=1***
*Else if word is noun and present in nouns hash set Count {word}++;*
***//Opinion Word Extraction***
***Step 7: for each sentence in the review database***
***Step 8: if (it contains a frequent feature, extract all the adjective words as opinion words)***
***Step 9: for each feature in the sentence***
***Step 10:*** *the nearby adjective is recorded as its opinion*
***//Opinion Orientation***
***Step 11:*** *Put* ***Opinion words*** *in OpinionHash*
***Step12:*** *Read files Positive.txt and Negative .txt*
***Step 13:*** *opinionsHash.put(line, 1)*
***Step14:****opinionsHash.put(line, -1)*
***//Summary Generation***
***Step 16:*** *Extract all* ***features*** *and their* ***opwords*** *into hash table*
***Step 17:*** *initialize NumOccurence to zero*
***Step 18: if OpHash contains opinion word then increment NumOccurence of that word***
***Else set num occurrence to 1***
***Step 19:*** *Output feature +opword+num occurrence +orientation*
***//Over all Summary Generation***
***Step 20: Orientation*** *ow =opHash.get (opword);*
***Step 21 : If (ow.orientation>0)***
*num pos ++;*
*num neg ++;*
***Step22:*** *Output feature + no of pos opinions and no of negative opinions.*

### A. Over all summary generation:

After all the previous steps, we are ready to generate the final feature-based review summary.

### a. Example of a Movie Review:

Probably yes, acting was good, plot brilliant, I mean Stephen King wrote the book, but it can't match the film. When I sow this movie I thought "that's it, this is it, best movie ever". Sure you can find people who didn't like it, and when I ask "why", and try to guess what then happened? They just confused themselves; they can say reason why they don't like this movie... When I try to read comments with 1/10star, I can't find comment with good reasons that could make me realize, that this movie is a bad movie, If there is no reason, then this means, that it's the best movie ever made! The only reason people are giving 1 star is that they don't like ratio "9, 2", they want to do more effect by pull this high ratio down, I really hate them :) Sorry for my bad English language, I live in Europe, my English teacher is big cow with no knowledge.

### b. Over all Summary:

Movie (Feature)
a) english : 4 : 0
b) good : 8 : 1
c) high : 5 : 0
d) best : 6 : 1
e) bad : 8 : -1

Reason (Feature)
a) What then : 1 : 0

Movie
a) Num Positive Opinions = 2
b) Num Negative Opinions = 3

Reason
a) Num Positive Opinions = 0

© 2010, IJARCS All Rights Reserved

b) Num Negative Opinions = 1

## V. EXPERIMENTAL RESULTS

*F*eature-*B*ased *S*ummarization based on the proposed techniques has been implemented in JAVA. We conducted our experiments using the user reviews of five movies. The reviews were collected from IMDb.com. Movies in these sites have a large number of reviews.

Table 1 : Experimental Results

| Movie Review | Opinion Sentence Extraction | | Sentence Orientation |
|---|---|---|---|
| | Recall | Precision | Accuracy |
| Movie1 Review | 0.719 | 0.643 | 0.927 |
| Movie2 Review | 0.634 | 0.554 | 0.946 |
| Movie3 Review | 0.675 | 0.815 | 0.764 |
| Movie4 Review | 0.784 | 0.589 | 0.842 |
| Movie5 Review | 0.653 | 0.607 | 0.730 |
| Average | 0.693 | 0.642 | 0.842 |

**Table 1** shows the evaluation results of the two procedures: opinion sentence extraction and sentence orientation prediction. The average recall of opinion sentence extraction is nearly 70%. The average precision of opinion sentence extraction is 64%.

## VI. CONCLUSION

In this paper, we proposed a set of techniques for mining and summarizing movie reviews based on data mining and natural language processing methods. The objective is to provide a feature-based summary of a large number of user reviews for movies. Our experimental results indicate that the proposed techniques are very promising in performing their tasks.

## VII. REFERENCES

[1] Bing Liu.Sentiment analysis and Opinion Mining April22nd 2012 Research on Opinion mining *www.cs.umn.edu/~blu.*

[2] Morinaga, Satoshi, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. *Mining product reputations on the web*. in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*. 2002.

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

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

[5] Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*, MIT Press.Cambridge, MA: May 1999.

[6] Minqing Hu and Bing Liu.Mining. Opinion Features in customer reviews. http://citeseerx.ist.psu.edu

### Short Bio Data for the Authors

**J.Kanaka Priya** received B-Tech degree in Computer Science & Engineering from Gayatri Vidya Parishad college of Engineering,Affiliated with JNTU-K in 2011 and presently pursuing M-Tech (2011-2013) in Gayathri Vidya Parishad college of Engineering (Autonomous), Visakhapatnam, INDIA. My areas of research include Image Processing.

**V.Adilakshmi** presently pursuing M-Tech (2011-2013) in Gayatri Vidya Parishad college of Engineering (Autonomous), Visakhapatnam, INDIA. My areas of research include Image Processing

**G.V.Hindumathi** recieved M.Tech in the field of Computer Science from Andhra University, and I had completed my B.Tech in department of CSE from GITAM and working as Asst Professor in department of Computer Science. My areas of research include Networks and Data mining.

© 2010, IJARCS All Rights Reserved