International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-10)

Research Article

October
2016

Sentiment Orientation of Smiley Based Comments for a Movie Review: Naïve Approach

Deenath Kumar G, Dr. J. Meenakumari
Department of Computer Science and Applications, The Oxford College of Science,
Bangalore, India

Abstract-

There is clear evidence from the literature review that lot of research has happened in the area of sentiment analysis. However, all these research studies were pertaining to text type. Very few works has combined text along with basic emoticons. These emoticons are generated with the help of special symbols in keyboard and they help to classify the reviews faster. A picture is worth thousand words in this context smiley based reviews help to understand the context better and with ease. Hence in this paper reviews of a specific movie are analyzed by several steps of opinion mining such as data cleaning, data integration, data selection, data transformation etc. The opinions (reviews) are considered to be a mixture of text and a smiley's. The set of opinions are gathered and integrated for analysis.

Keyword- Sentiment Analysis; Smiley Based Review Mining; Lexicon based analysis;

I. INTRODUCTION

People are habitual and inquisitive to know what others feel or what other opinion on something is before making a decision by themselves. With the availability and popularity of opinion-rich resources such as online review sites and personal blogs too, new opportunities and challenges arise as people can, and do actively use information technologies to seek out and understand the opinion of others. The abrupt eruption of an activity in the area of sentiment analysis, which deals with the computational treatment of an opinion, sentiment and subjectivity in text, has thus occurred at least in part as a direct response to the delve of an interest in the new systems that deals directly with an opinion as first class object.

To help out audience in making decision towards the subject, it is necessary to understand the reviews on a particular subject based on the keywords in the reviews.

The thesis focuses on analyzing the rating for a movie automatically without detecting Aspect Keyword. To analyze the review, text mining and Sentiment Analysis approach is used. The purpose of Text Mining is to process unstructured information (textual), exact meaningful indices from the text and thus make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. And in the work reviews are considered not only as text but along with smileys. Sentiment Analysis refers to the use of "beyond polarity" Sentiment classification looks, for instance at emotional states such as "good", "poor" and "satisfied". The algorithm proposed for this work is Naïve-Bayes classification.

In the work, keywords are manually considered from AFFIN List; is a list of English words rated for valence with an integer between minus five (negative) to plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated. There are two versions: AFINN-111: Newest version with 2477 words and phrases. AFINN-96: 1468 unique words and phrases on 1480 lines. The Naïve-Bayes algorithm detects the affinity values for each keyword defined in AFFIN List which is ranging from -5 to +5. The values for entire keywords are summed up to acquire the rating for movie. As the rating for the product is obtained, a tagline along with smiley is given which describes the movie is satisfying the audience or not which clears the audience confusion before watching a particular movie.

The dataset for the thesis is collected from Review Entertainer (*www.reviewentertainer.com*). The dataset is a collection of data related to movies of various languages. Any audience will have a benchmark or expectation before watching a movie. So before making a decision about a movie audience will consider various parameters about a movie such as direction, music, story, screenplay and action etc. Hence based these factors users or critics write their reviews for a movie. Theses reviews are processed to give the exact rating for a movie using the proposed algorithm.

II. RELATED WORK

Historically, sentiment analysis has been concerned with assigning a binary classification to sentences or entire documents that represents the polarity (i.e., the orientation) of the writer towards the discussed contents [12, 13]. Nevertheless, the overall polarity gives no indication about which aspects the opinions refer to. For this reason, in 2004 Hu and Liu [14] introduced the more interesting problem of aspect- based sentiment analysis, where polarity is not assigned to sentences or documents, but to single aspects discussed in them. In their approach, given a large number of reviews for a specific product, they first attempt to identify.

Qiu et al. [15] continued to pursue the idea that opinion words can be used to detect product aspects and vice versa, focusing on single reviews. In their approach, a seed set of opinion words is combined with syntactic dependencies to identify product aspects and new opinion words. To detect the polarity of the newly identified opinion words, they consider the given polarities of the seed words and make the assumption that opinion words expressing a sentiment towards the same aspect in the same review share the same polarity. While Qiu et al. use syntactic dependencies solely to capture word sequences that contain aspects or opinion words already observed, our approach uses dependency paths to detect new product aspects, with the potential advantage of achieving higher coverage.

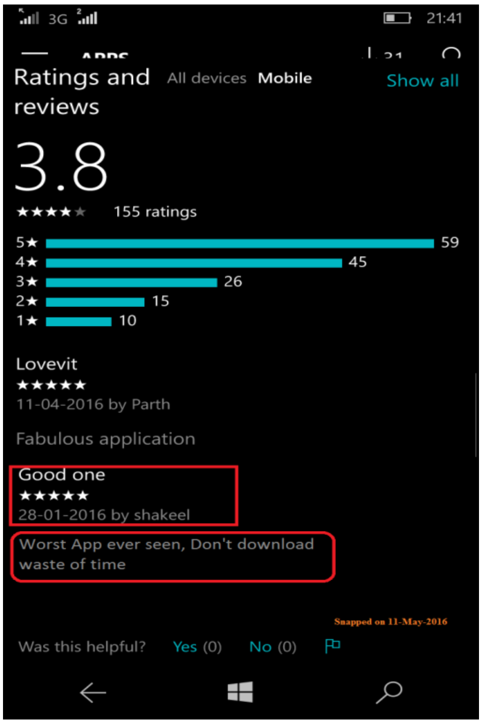
A different line of work on aspect-based sentiment analysis is based on topic models. Brody and Elhadad [16] have tried to use Latent Dirichlet Allocation (LDA) [17] to extract topics as product aspects. To determine the polarity towards each topic/aspect, they start from a set of seed opinion words and propagate their polarities to other adjectives by using a label propagation algorithm. Instead of treating aspect detection and sentiment classification as two separate problems, Lin and He [18] and Jo and Oh [19] directly integrate the sentiment classification in the LDA model, so that it natively captures the sentiment towards the topic/aspect.

In this research, we study the problem of generating *feature-based summaries* of audience reviews of the movies. Here, *features broadly* mean product features (or attributes) and emotions associated with a particular movie. Given a set of audience reviews of a particular movie, the task involves three subtasks: (1) identifying features of the product that audience have expressed their opinions on (called *movie comments*); (2) for each feature, identifying review sentences that give positive, negative or neutral opinions; and (3) producing a summary using the discovered information [6].

III. PROBLEM STATEMENT

In the era of Web 2.0, more and more people express their opinions on all kinds of entities, including products and services, which in turn help not only customers make informed decisions but also merchants improve their services. A main challenge in SA is attaining exact rating for a gven subject while user is provided with an interface to rate (not review) corresponding subject. Say for example user is given interface of both rate and review. Here in this case he or she may rate a review they wish but there are some chances that user could enter negative review and rate the product with high scale; inturn will affect the overall rating to wrong.

Same case can be observed in Windows Mobile Appliction Store for rating 'Ringtone Maker' application as shown in Fig 3.1.



Ratings and reviews

3.8

★★★★★ 155 ratings

Lovevit
★★★★★
11-04-2016 by Parth

Fabulous application

Good one
★★★★★
28-01-2016 by shakeel

Worst App ever seen, Don't download waste of time

Snapped on 11-May-2016

Fig 3.1 Invalid review and rating of Ringtone Maker App – Picture from Windows store on 11-May-2016

In the above figure user has rated with almost high scale by giving negative comment. Since user rated with high magnitude the commet is highlighted as 'Good one'. And also affected overall rating of that application.

IV. OBJECTIVE AND METHODOLOGY

4.1 Objective

The objective of the work is to automatically analyzing the rating a movie. The dataset considered is a collection of data related to movies. When audience comes online to check for rating for a movie various parameters he or she would consider which are direction, screenplay, actor performance etc. Based on these parameters users or critics write their reviews for the product. These reviews are processed to give the exact rating for a movie using the proposed algorithm Naïve Bayes Classification. For the dataset containing data related movie reviews the each keyword is refereed with AFFIN List to get the equivalent polarities of keywords.

4.2 Methodology

A sentiment orientation is carried out on dataset of various language movie reviews taken from reviewentertainer.com and data is cleaned and transformed into favorable conditions of system. A challenge in the work is when the huge number of reviews for a particular movie like is given the system has to split each keyword on transformed data. To solve this challenge, Naïve Bayes Classifier Algorithm for Smiley Based Opinion model is proposed in order to discover the values for each keyword and the total rank for the entire review. Thus, our overall approach consists of two stages, which we will further discuss in detail.

4.2.1 Naïve Bayes Algorithm for Smiley Based Reviews

In the initial phase input reviews {d} and the review considered is subdivided into subsets, X. Matching keywords are found in the AFFIN List, then the polarity values of keywords are rendered through AFFIN List and stored in count variable. This process is looped till the last review is encountered. While looping, each sentence is oriented as any of these four categories which are Very Negative, Negative, Positive or Very Positive. On finding the maximum among counter variables of each category the sentence is oriented the Summing up all these word values gives rating for that particular aspect word. This flow is repeated for all the aspect keywords and the rating for the overall review is generated.

4.2.3 Algorithm

Input: A review $\{d\}$, Set of aspect keywords for a movie $\{T1, T2, \ldots\}$,
Output: Factors (+,-,|) for loaded reviews in percentage along with overall rating of loaded reviews. AffinList V, selection threshold p and iteration step limit I.
Step 1: Review d divided into subsets, $X = \{x1, x2... xM\}$;
Step 2: Match the keywords in each subset of X and sum up each score of word retrieved from AFFIN List (-5 to +5);
Step 3: Calculate number of matches from -5 to +5 of each keyword for at least p review (Very Negative to Very Positive).
Step 4: Find out mVal = max (VN, N, P, VP)
Step 5: +: VP ≥ mVal ≥ P; -: VN ≥ mVal ≥ N; |: (VN+N+P+VP) ==0
Step 6: Orient each review as in Step 5 and hold count C for each factors i.e.., +, -, |
Step 7: Output the C (in %) for +, - and | and Reduce + in scale of 10 to get overall rating using the below formula.

$$\text{Rate} = \left[\frac{(+_c \,/\, \text{sum}\ (+_c,\ -_c,\ |_c))}{(-_c \,/\, \text{sum}\ (+_c,\ -_c,\ |_c))}\right] \div 10$$

Where $+_c$ Positive Count for n number of reviews loaded
$-_c$ Negative Count for n number of reviews loaded and
$|_c$ Neutral Count for n number of reviews loaded.

4.2.4 The Process

Specially, the basic workflow of the proposed Naïve Bayes Classifier for Smiley based Opinions is follows: The review {d} based on the movie review User can input n number of reviews to load from the dataset. From the reviews {d} obtained; also the smiley's are converted to Unicode and these Unicode to mapped to get the appropriate emotion keywords. Selection threshold p and iteration step limit I are considered as inputs for the algorithm.

The input review {d} is divided to form subsets nothing but divide the reviews in to set of sentences $X = \{x1, x2... xM\}$. From these subsets special characters, unwanted symbols and spaces are deleted. The matching keyword in the subset X are discovered and mapped with the AFFIN List to get polarity value for that keyword and store in temporary variable. After matching each keyword in a sentence the polarity value is summed up and simultaneously maintaining the count variable for each category for a sentence which as Very Negative (VN), Negative (N), Positive (P) and Very Positive (P).

This procedure is carried out until no review is left behind i.e., the above procedure is looped till the n^{th} review. And to orient a movie maximum value (mVal) of VN, N, P, VP is determined. If the value mVal is lying between VP and P then a review is treated as positive (+), similarly, a review is oriented as negative (-) if the mVal is lying between VN and N. To check for neutral reviews we are summing up the count variable for VP, P, N, and VN. If the obtained sum is zero then a review is neutral. Finally, each category (+, - and |) for loaded number of reviews are converted into percentage (%) for understanding purpose. To get overall rating for a loaded n number of movie review(s) the international movie database standards is maintained in the work, where the overall rating is made in the scale of 10. The overall rating for a movie is calculated using the formula

$$\text{Rate} = \left[\frac{(+_c \,/\, \text{sum}\ (+_c,\ -_c,\ |_c))}{(-_c \,/\, \text{sum}\ (+_c,\ -_c,\ |_c))}\right] \div 10$$

V. ANALYSIS

5.1 Dataset and Preprocessing

The dataset is considered is a collection of movie reviews related to various languages.
Preprocessing steps are followed in the review collected

1) Read the review collected; 2) Review is divided into subsets and corpus occurrence; 3) Converting words into lower cases; 4) removing punctuations, stop words, stripping whitespaces.

Here in the following Fig 5.1 is an example sample review for an English movie 'The Mask' which contains punctuations, stop words. This has to be considered for preprocessing.

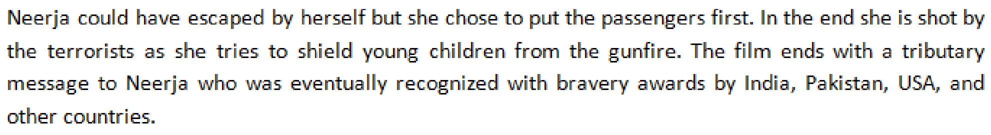

Neerja could have escaped by herself but she chose to put the passengers first. In the end she is shot by the terrorists as she tries to shield young children from the gunfire. The film ends with a tributary message to Neerja who was eventually recognized with bravery awards by India, Pakistan, USA, and other countries.

Fig. 5.1 Sample Review before Pre-processing

The Fig 5.2 shows the same review after preprocessing; removing stop words, punctuations and all the characters in the review are converted to lower case.

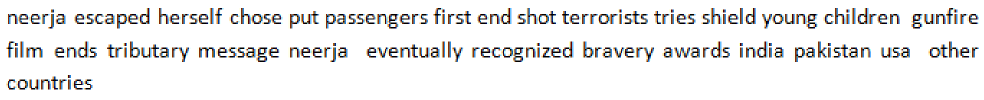

neerja escaped herself chose put passengers first end shot terrorists tries shield young children gunfire film ends tributary message neerja eventually recognized bravery awards india pakistan usa other countries

Fig. 5.2 Sample Review after Preprocessing

5.2 Smiley Emotion Identification

In the work there are 35 types of smiley are used for reviews in turn 35 Unicode could be mapped with reference to AFFIN List. Sometimes same type of smiley's are used for reviewing a movie for example consider a smiley, 😋 the Unicode for this smiley is U1F60B, when we match this smiley into normal English word it will be 'delicious' but on the same time it can be 'crazy' too. So when we refer a keyword either delicious or crazy the polarity value for both the keywords will be same.

5.3 Sentiment Score Calculation

Once the reviews are cleaned and transformed into system favorable conditions the sentiment score will be calculated using counter variables for each category which are Positive Count $(+_c)$, Negative Count $(-_c)$ and $(|_c)$Neutral Count is converted into percentage.

Fig. 5.3 shows sentiment score obtained for a Kannada movie Ulidavaru Kandante (As seen by the rest) on loading 100 reviews.

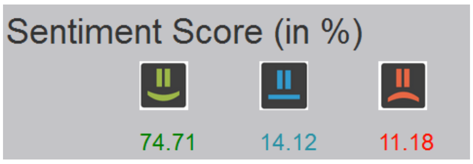


Fig 5.3 Sentiment Score for Kannada Movie Ulidavaru Kandante (As Seen by the Rest)

The system simultaneously will plot a 3DScatterPlot for Positive, Negative and Neutral Factors depending on the number of reviews loaded. Fig 5.4 shows a 3DScatterPlot for above experiment.

Similarly the above experiment is continued to get 2D Scatter Plot for Positive and Negative factors where Fig. 5.5 (a) plotted against Positive Factor Vs Negative Factor and performed Average Negative Factor which average possibility of getting negative factors on loading certain number of reviews.

Fig. 5.5 (b) plotted against Positive Factor Vs Negative Factor and performed Linear Regression for Negative Factor on Positive Factor which shows the amount of dependency between positive factor and negative factor.

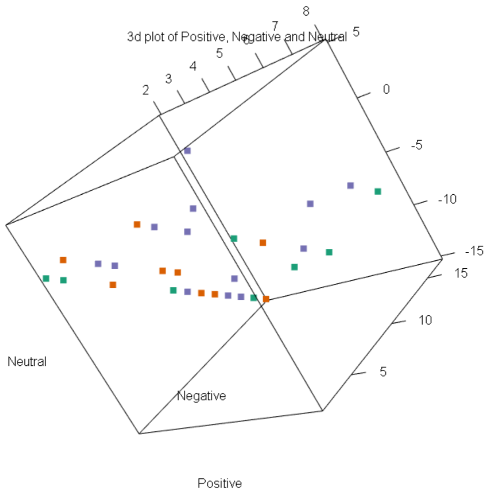


Fig 5.4 3D-ScatterPlot for Kannada Movie Ulidavaru Kandante (As Seen by the Rest)

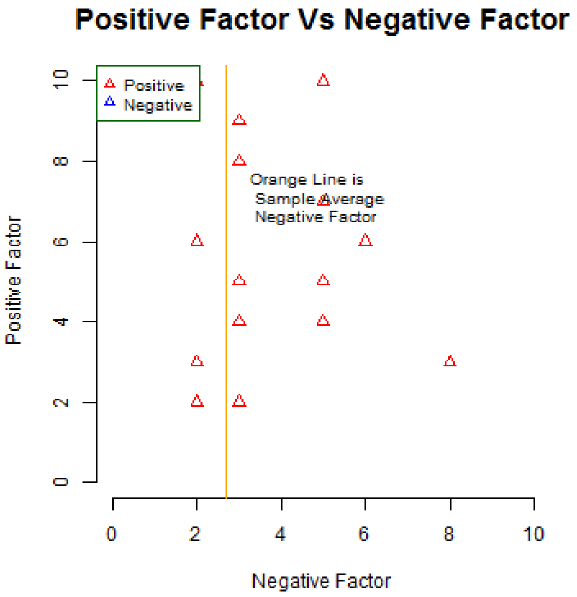


Fig 5.5 (a) Scatter Plot of Average Negative Factor for Kannada Movie Ulidavaru Kandante (As Seen by the Rest)

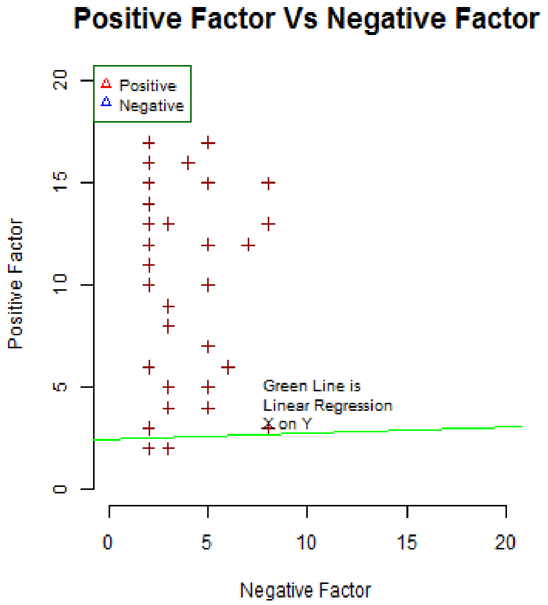


Fig 5.5 (b) Scatter Plot of Linear Regression for a Kannada Movie Ulidavaru Kandante (As Seen by the Rest)

Similarly, for the above experiment the wordcloud is also plotted and shown below.

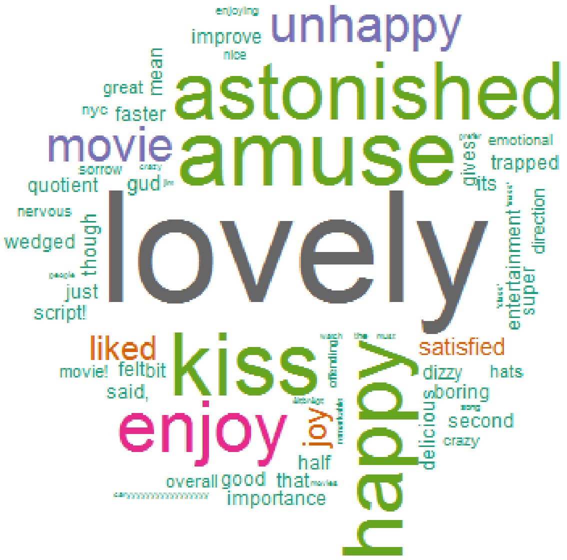
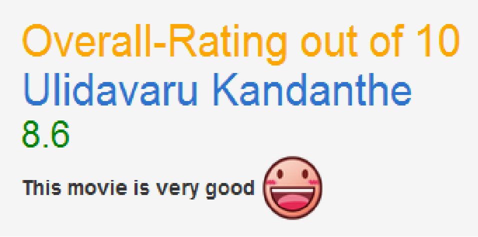


Fig 5.6 WordCloud for Kannada Movie Ulidavaru Kandante (As Seen by the Rest)

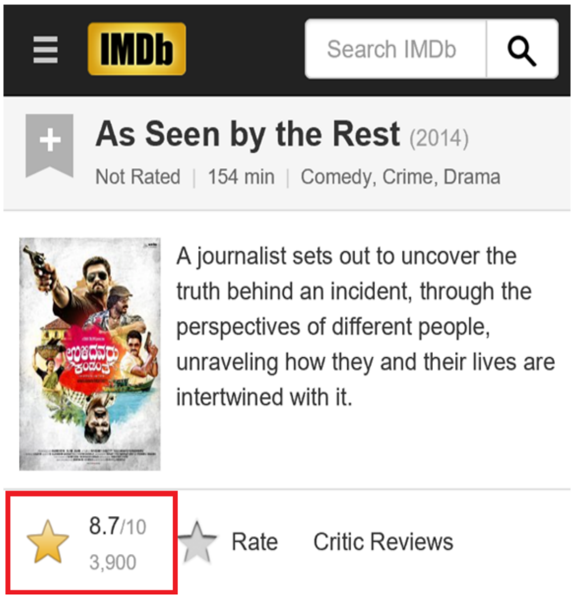
Finally on calculating the overall rating and comment on movie is shown below in Fig5.11.



Overall-Rating out of 10
Ulidavaru Kandanthe
8.6
This movie is very good

Fig. 5.11 Overall Rating for Kannada Movie Ulidavru Kandante (As Seen by the Rest)

I also compared the overall rating for a movie Ulidavaru Kandanthe with IMDb; and it is found that the obtained result using Naïve Bayes Classifier for Smiley Based Opinion is achieving accuracy ≈ 98.85%.



As Seen by the Rest (2014)
Not Rated | 154 min | Comedy, Crime, Drama

A journalist sets out to uncover the truth behind an incident, through the perspectives of different people, unraveling how they and their lives are intertwined with it.

8.7/10
3,900

Fig 5.12 International Movie Database (IMDb) rating for Kannada Movie Ulidavaru Kandanthe (As Seen by the Rest)

VI. FUTURE ENHANCEMENT

In the work "Sentiment Orientation of Smiley Based Opinion with Keywords for the Movie Reviews" focuses on analyzing the overall rating for a movie automatically. The dataset considered is movie reviews. In future, the work can be extended to other domains with different set of aspect keywords and also for movie review dataset the other parameters defining feature of a movie can also be considered. The work can be extended to get overall rating for each aspects of a movie. And the overall rating and auto commenting can be enhanced using any hybrid sentimental analysis algorithm.

VII. CONCLUSION

In this work, a new unified generative Naïve Bayes Classifier for Smiley Based Reviews is proposed and tested. A Naïve Bayes Classifier explores the methodology to extract the exact overall rating for a movie based on the keywords used in the review written by users/audience for the dataset of movie reviews of various languages. In addition an overall rating which describes whether a movie satisfies audience expectations or not is also provided.

REFERENCES

[1] Deenath Kumar G., *"A Survey on Sentiment Analysis Types & Approaches: Overview"*, CTCS-16, pages 94 - 97, ISBN 978-93-85477-77-5 Mar. 2016

[2] Chakrabarti. S., "Mining the Web: Discovering Knowledge from Hypertext Data". Morgan Kaufmann, 2002.

[3] Han,J. and Chang, K. C.-C.,"*Data Mining for Web Intelligence*", IEEE Computer, Nov. 2002.

[4] Lan Yi, Bing Liu, Xiaoli Li., "*Eliminating Noisy Information in Web Pages for Data Mining*", School of Computing National University of Singapore 3 Science Drive 2 Singapore 117543, 2003.

[5] Sara Owsley Sood, Elizabeth F. Churchill, Judd Antin., "*Automatic identification of personal insults on social news sites"*, Pomona College 185 East Sixth Street Claremont, CA 91711, 2009.

[6] Minqing Hu and Bing Liu., "*Mining and Summarizing Customer Reviews*", Minqing Hu and Bing Liu Department of Computer Science University of Illinois at Chicago 851 South Morgan Street Chicago, IL 60607-7053, 2004.

[7] X. Ding, B. Liu, and L. Zhang., "*Entity discovery and assignment for opinion mining applications"*, In Proceedings of the 15th KDD, pages 1125–1134. ACM, 2009.

[8] Y. Lu and C. Zhai., "*Opinion integration through semi-supervised topic modeling*", In Proceeding of the 17th WWW, pages 121–130. ACM, 2008.

[9] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima., *"Mining product reputations on the web*", In Proceeding of the 8th KDD, pages 341–349, 2002.

[10] M. Hu and B. Liu., "*Mining and summarizing customer reviews*", In Proceedings of the 10th KDD, pages 168–177. ACM, 2004.

[11] N. Jindal and B. Liu., "*Identifying comparative sentences in text documents"*, In Proceedings of 29th SIGIR, 2006.

[12] Paltoglou G., Thelwall M., *"A study of information retrieval weighting schemes for sentiment analysis"*, In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1386–1395, 2010.

[13] Yessenalina A., Yue Y., Cardie C., *"Multi-level structured models for document level sentiment classification"*, In Proceedings of the 2010 conference on Empirical Methods in Natural Language Processing, pages 1046–1056, 2010.

[14] Hu M., Liu B., *"Mining and summarizing customer reviews"*, In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 168–177. ACM, 2004.

[15] Qiu G., Liu B., Bu J., Chen C., *"Expanding domain sentiment lexicon through double propagation"*, In Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence, volume 9, pages 1199–1204, 2009.

[16] Brody S., Elhadad N., "*An unsupervised aspect-sentiment model for online reviews"*, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 804–812. Association for Computational Linguistics, 2010.

[17] Blei D. M., Ng A. Y., Jordan M. I., *"Latent dirichlet allocation"*, Journal of Machine Learning research, 3:993–1022, 2003.

[18] Lin C., He Y., *"Joint sentiment/topic model for sentiment analysis"*, In Proceedings of the 18th ACM Conference on Information and Knowledge Management, pages 375–384, 2009.

[19] Jo Y., Oh A. H., *"Aspect and sentiment unification model for online review analysis"*, In Proceedings of the fourth ACM international conference on Web search and data mining, pages 815–824, ACM, 2011.

[20] Tait J., "*AutomaticSummarizing of English Texts"*, Ph.D. Dissertation, University of Cambridge, 1983.

[21] Hu, M., and Liu, B., *"Mining Opinion Features in Customer Reviews*", to appear in AAAI'04, 2004.

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

[22] M. Hu and B. Liu., *"Mining opinion features in customer reviews"*, In AAAI, pages 755–760. AAAI Press / The MIT Press, 2004.

[23] I. Titov and R. T. McDonald., *"A joint model of text and aspect ratings for sentiment summarization"*, In Proceedings of the 46th ACL, pages 308–316, 2008.

[24] L. Zhuang, F. Jing and X.-Y. Zhu., *"Movie review mining and summarization"*, In Proceedings of the 15th ACM international conference on Information and knowledge management, 2006.

[25] Wouter Bancken, Daniele Alfarone and Jesse Davis., "*Automatically Detecting and Rating Product Aspects from Textual Customer Reviews"*, Department of Computer Science, KU Leuven Celestijnenlaan 200A - box 2402, 3001 Leuven, Belgium, 2014.

[26] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty., *"Latent dirichlet allocation. Journal of Machine Learning Research"*, 3:993–1022, 2003.

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

[28] Morinaga, S., Ya Yamanishi, K., Tateishi, K, and Fukushima, T., *"Mining Product Reputations on the Web"*, KDD 2002.

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

[30] Cooley, R., Mobasher, B. and Srivastava, J., *"Data preparation for mining World Wide Web browsing patterns"*, Journal of Knowledge and Information Systems, (1) 1, 1999.

[31] Davision, B.D., *"Recognizing Nepotistic links on the Web*", Proceeding of AAAI. 2000.

[32] Hongning Wang, Yue Lu, Chengxiang Zhai., "Latent Aspect Rating Analysis without Aspect Keyword Supervision", Department of Computer Science University of Illinois at Urbana-Champaign IL, 60801 USA, 2011.

[33] Shian-Hua Lin and Jan-Ming Ho., "Discovering Informative Content Blocks from Web Documents", KDD 2002.

[34] Lee, M.L., Ling, W. And Low, W.L., "Intelliclean: A knowledge-based intelligent data cleaner", KDD-2000.

[35] Nahm, U.Y., Bilenko, M. and Mooney R.J., *"Two Approaches to Handling Noisy Variation in Text Mining"*, ICML-2002 Workshop on Text Learning, 2002.

[36] Jushmerick, N., "Learning to remove Internet advertisements", AGENT-1999.

[37] Yang, Y. and Pedersen, J.O., "A comparative study on feature selection in text categorization", ICML 1997

[38] Kao, J.Y., Lin, S.H. Ho, J.M. and Chen, M.S., "Entropy-based link analysis for mining web informative structures", CIKM 2002.

[39] Kleinberg, J., *"Authoritative Sources in a Hyperlinked Environment*", ACM- SIAM Symposium on Discrete Algorithms, 1998.

[40] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai., *"Topic sentiment mixture: modeling facets and opinions in weblogs"*, In Proceedings of the 16th WWW, pages 171–180. ACM, 2007.

[41] C. Lin and Y. He., *"Joint sentiment/topic model for sentiment analysis"*, In Proceeding of the 18th CIKM, pages 375–384, New York, NY, USA, 2009. ACM.

[42] Y. Jo and A. H. Oh., *"Aspect and sentiment unification model for online review analysis"*, In Proceedings of the fourth ACM international conference on Web search and data mining, 2011.