Global Journal of Computer Science and Technology: G Interdisciplinary
Volume 15 Issue 2 Version 1.0 Year 2015
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Detecting Sentiments from Movie Reviews by Integrating Reviewer's Own Prejudice

By Kalpana Yadav, Sumit K. Yadav & Swati Gupta
Indira Gandhi Delhi Technical University, India

Abstract- Presently, sentiment analysis algorithms are widely used to extract positive or negative feedback scores of various objects on the basis of the text/reviews. But, an individual may have a certain degree of biasness towards a certain product/company and hence may not objectively review the object. We try to combat this biasness problem by incorporating the positive and negative bias component in the existing sentiment score of the object. This paper proposes several algorithms for a new system of implementing individual bias in the corpus of data i.e. movie reviews in this case. Each review comment has an unadjusted sentiment score associated with it. This unadjusted score is refined to give an adjusted score using the positive and negative bias score. The bias score is calculated using certain parameters, the weightage of which has been determined by conducting a survey. We lay emphasis on the degree of biasness an individual has towards or against the review parameters for the movie reviews corpus namely actor, director and genre. We equip the system with the capability to handle various scenarios like positive inclination of the user, negative inclination of the user, presence of both positive and negative inclination of the user and neutral attitude of the user by implementing the formulae we developed.

Keywords: natural language processing, sentiment analysis, opinion mining, text classification, online customer reviews, social network analysis.

GJCST-G Classification: I.3.3

Detecting Sentiments from Movie Reviews by Integrating Reviewer's Own Prejudice

Kalpana Yadav [α], Sumit K. Yadav [σ] & Swati Gupta [ρ]

Abstract- Presently, sentiment analysis algorithms are widely used to extract positive or negative feedback scores of various objects on the basis of the text/reviews. But, an individual may have a certain degree of biasness towards a certain product/company and hence may not objectively review the object. We try to combat this biasness problem by incorporating the positive and negative bias component in the existing sentiment score of the object. This paper proposes several algorithms for a new system of implementing individual bias in the corpus of data i.e. movie reviews in this case. Each review comment has an unadjusted sentiment score associated with it. This unadjusted score is refined to give an adjusted score using the positive and negative bias score. The bias score is calculated using certain parameters, the weightage of which has been determined by conducting a survey. We lay emphasis on the degree of biasness an individual has towards or against the review parameters for the movie reviews corpus namely actor, director and genre. We equip the system with the capability to handle various scenarios like positive inclination of the user, negative inclination of the user, presence of both positive and negative inclination of the user and neutral attitude of the user by implementing the formulae we developed. Hence, the system computes an objective score sans any individual bias for several scenarios making inferences better.

Keywords: natural language processing, sentiment analysis, opinion mining, text classification, online customer reviews, social network analysis.

I. Introduction

Sentiment analysis or opinion mining is the field of natural language processing dedicated to the computational analysis of opinions for the purpose of decision making (Kim, & Hovy, 2004). An opinion is a statement about a subject which expresses the sentiments and emotions of the opinion maker on the subject.

The main objective of sentiment analysis is to extract relevant information about the various sentiments articulated by authors about a particular subject, forming relationship patterns between the sentiments and the subject and helping users by presenting the huge volume of unstructured Web data in a structured form. (Wu, Wang, & Yi, 2013). In the present Internet age there is a plethora of information available to the users in every possible arena. Users are exposed to various sources of information like blogs, online reviews, and social sites.

The current trend is to look up reviews, expert opinions and discussions on the Web, so that one can make an informed decision pertaining to day-to-day tasks and purchases. (Cui, Mittal, & Datar, 2003). With so much information around, the user finds it difficult to process all of it and make an informed and rational decision. Here, sentiment analysis plays an important role by analysing all the data available and providing an over-all positive or negative feedback. (K. Dave, S. Lawrence, & D. Pennock, 2003).

Presently, Internet is extensively used as a platform for shaping up (Zuniga, Puig-I-Abril, & Rojas, 2009) views of people in diverse fields like politics (Park, Ko, Kim, Liu, & Song, 2011), (Larsson, & Moe, 2011) religious ideology, business marketing, tourism (Claster, Cooper, & Sallis, 2010) book reviews(Lin, Fang, & Wang, 2013) etc. Hence, it becomes imperative to have a mechanism to sift through this prejudiced information and get a collective objective consensus on the whole. For this evaluation, the validity of the source becomes equally important along with the content expressed.

The content authors can be classified into three types: promoters, the users who are positively prejudiced towards the object; detractors, the users who are negatively prejudiced against the object and passives, the users who are neither positively nor negatively inclined towards the object. (Wen, Dai, & Zhao, 2012). The bias or prejudice mentioned above refers to the inclination of temperament to hold a partial perspective and a refusal to even consider the possible merits of alternate points of view. The different forms of bias that have already been explored in sentiment analysis field include herd behavior, (Chen, 2008) first impression bias, (Deffuant, & Huet, 2009) sequential bias, (Piramuthu, Kapoor, Zhou, & Mauw, 2012).

The system we propose aims to deal with the individual bias in order to evaluate the validity of the content sources and hence get an objective consensus rather than the subjective (Liu, 2010) one that we are previously exposed to. Our work focuses on movie

Author α : Kalpana Yadav, Assistant professor in Information Technology Department at Indira Gandhi Delhi Technical University, Delhi. e-mail: kyadav11@yahoo.com

Author σ : Sumit K Yadav, Assistant professor in Computer Science Department at Indira Gandhi Delhi Technical University, Delhi. e-mail: sumitarya007@gmail.com

Author ρ : Swati Gupta, Student in Computer Science Department at Indira Gandhi Delhi Technical University, Delhi. e-mail: gupta.swati1992@gmail.com

reviews corpus dataset as it provides a wholesome sample data from varied demographics since movies are watched by everyone.

II. Related Work

In this section, we focus on the related work on various types of bias and sentiment classification especially in the field of online reviews.

a) *Sentiment Classification*

The sentiment analysis has evolved over the period of time be it examining semantic orientation (Hatzivassiloglou, & McKeown, 1997) of adjectives, of adverbs,(Turney, & Littman, 2003) of emoticons (M, A. K. K., 2011) of different languages, (Martín-Valdivia, Martínez-Cámara, Perea-Ortega, & Ureña-López, 2013) of compound sentences using sentence level analysis (Mishra, & Jha, 2013), usage of appraisal groups (Whitelaw, Garg, & Argamon, 2005). and unsupervised techniques. The granularity of data mining has also evolved from document level, (Pang, Lillian, & Shivakumar, 2002) (Turney, 2002) sentence level (Riloff, & Janyce, 2006) to object level (Hu, Minqing, & Liu, 2004) techniques.

In (Cui, Mittal, & Datar, 2003), the efficiency of high order n-grams is enhanced using discriminating classifier. Also, the possibility of getting a consolidated result even with the data set comprising of varied products and authors is explored in this paper. (Dou, & Hu, 2012) explores an automated method incorporating semantic analysis and align technique to extract structured data form web pages has been developed. (Huang, & Lin, 2013) has dealt with a system where product reviews are evaluated on three parameters: product reviews, product popularity, and product release month and a proficient product ranking system is created. In (Jusoh, & Alfawareh, 2013) the sentiment classification using possibility theory has been implemented in order to determine varied degree of positive and negative sentiment score.

b) *Bias Reviews*

Various types of bias have also been discovered in the papers. In (Bencz, A. 2012), bi-clustering has been used along with kernel methods and baseline text classifiers to improve trust, bias and factuality classification over Web data on the domain level. The main aim is to aid researchers in obtaining large data that originates from trustworthy sources.

In (Sikora, & Chauhan, 2012), the first impression bias i.e. the tendency of the individuals to modify their opinions on the basis of first- third person review that he/she views which has been eliminated using the Kalman filtering technique. In (Schweiger, Oeberst, & Cress,2014), the confirmation bias in web based search was studied. The two data samples taken were psychotherapy and pharmacotherapy both of which are scientifically equally effective for depression treatments but the former was considered to be more effective by the public. The blog entries by experts and tag clouds were recommended to counter biased information processing on these entries. In (Wood, & Dellarocas, 2006), the reporting bias of the traders and its effects on the public feedback have been studied. The basic assumption dealt with here is that the traders are more likely to report or give a feedback when the experience has been positive rather than when it is negative. Hence, the lack of negative feedback doesn't necessarily mean the absence of it. In (Hu, Bose, Gao, & Liu, 2011), a simple statistical method has been developed in order to detect the online product reviews which are biased and how they affect the consumer reaction to the products. The two parameters on which review manipulation were judged were manipulation through ratings and manipulation through sentiments. The consumers were found to have detected successfully only the former.

In (Piramuthu, Kapoor, Zhou, & Mauw, 2012) sequential bias in the online product of the recommender systems are found and eliminated. In (Sikora, & Liangjun, 2014) the various methods used by traders to alter their reputation score in the online market have been studied. Here, the concept of replicator dynamics is used to study the evolution of different types of sellers and buyers in the market. In (Chen, & Lin, 2013), decision tree along with correlation analysis and extracted knowledge rules has been used to improve the detection of the online review manipulation by introducing eight review manipulation attributes. In (Hu, Bose, Gao, & Liu, 2011) the study on the increase in propensity of biasness in the book reviews increases with the passage of time has been explored. In (Cipriani, Guarino, & Antonio,2012), the herd behavior in financial markets has been studied and eliminated using structural estimation framework.

The paper (Knight, & Chiang, 2008) investigates the media bias and the influence the media has on casting of votes during election time. The paper concludes that although newspapers do influence the opinion formation of the voters, it is limited by the degree and direction of the bias. In (Wang, Zhang, X. M., & Hann, 2010), the social bias in online product ratings has been explored. The degree of social influence was found to be greater for the books with that were popular, if the rating was from less experienced user, the rating was given at a later stage of review cycle and if the rating was given by a user with small social network.

After the literature review, we find that individual bias though mentioned in various papers has never been worked upon or researched on before. Since, individual bias is one aspect that can greatly modify the sentiment score, hence, we decided to concentrate on this topic as our area of work and present the user with an objective score.

III. Proposed Methodology

The proposed system has seven major steps which start from extraction of corpus for the formulae to be applied upon. The next step is to extract the user data which are the likes from his/her Facebook® profile and the profile URL and manage the database hence, created. This serves as input for the mathematical modeling of the system. The corpus extracted is fed to ALCHEMY API to give an unadjusted score for the corpus. Further, steps include mathematical modeling and application of the developed formulae to calculate adjusted and unadjusted score for the corpus. In the end, we present the user with an unadjusted score which is an objective score i.e. sans any individual bias. Framework of proposed approach is shown in Figure 1.

Step 1: Extraction of Movie Reviews (Sentimental Data) for Social Media

Movie reviews are collected from social media, weblogs, bloggers, social networking sites like Facebook®, Twitter etc. for further processing.

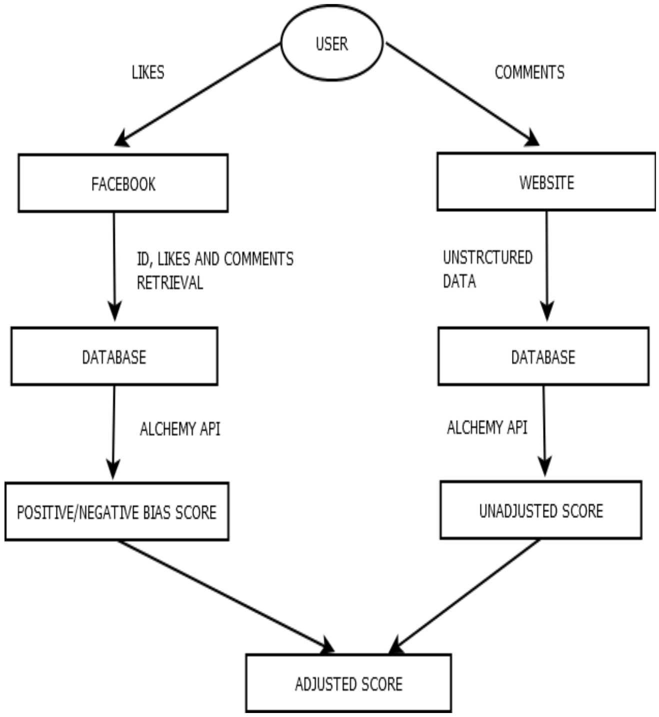

Figure1: Framework of the proposed methodology

Step 2 : Extraction of User Data

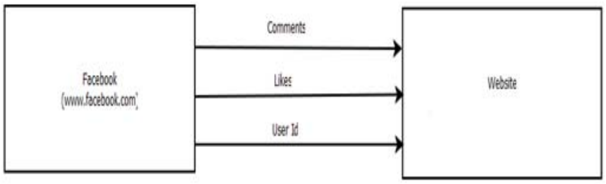

Figure 2 : Bidirectional Flow

In order to track the preferences and compute the likes and dislikes, the data related to the user i.e. the user's likes, their comments, the user ID, etc. is extracted from Facebook® in form of tokens. It is then sent to the website for further computation.

Step 3: Database Management

The comments, the likes, the user information and id are stored in *Phpmyadmin* database management system. The calculated sentiment score and bias score is also stored in the database. Next, unadjusted score is calculated using ALCHEMY API.

Step 4: Alchemy Api

The system makes use of a text analysis tool called ALCHEMY API (http://www.alchemyapi.com/). This tool provides the real-time text analysis through the method of entity and keyword extraction and provides the degree of positive and negative connotation they have.

It works on diverse document types including news articles, blog posts, product reviews, comments and Tweets.

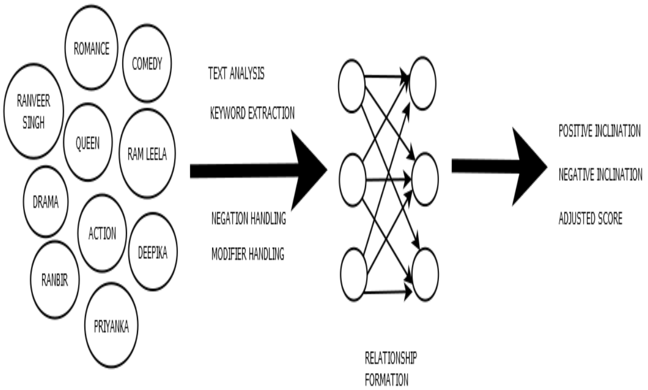

Figure 3 : Alchemy API Flow

The basic idea behind this framework is that it targets unstructured data, forms relationship between the keywords and the data and gives the relevant structured result. Figure 3 showcases the working of the ALCHEMY API.

The keywords in this figure are of three types: the keywords representing Bollywood actor/actress namely Ranbir, Deepika, Priyanka, Ranveer Singh; the keywords representing Bollywood movie names namely Ramleela and Queen; the keywords representing movie genre namely Action, Comedy, Drama and Romance.

The ALCHEMY API applies multiple algorithms of text analysis, keyword extraction, negation handling and modifier handling on these keywords and gives a structured relationship between them. The final result is in the form of positive and negative bias score.

Step 5: Mathematical Modeling

To determine the relevant parameters and their corresponding weightage to analyze the corpus a preference survey of the varied sample of a movie audience was conducted.

Thus, the movie reviews are analyzed on two major factors namely genre and actor/director of the movie in order to determine the bias of an individual.

a) Genre

The genre refers to the style or category of the movie for example Drama, Romance, Action, among others.

b) Actor/Director

The user inclination towards or against certain actors and director in the movie can make a user biased towards the movie as well.

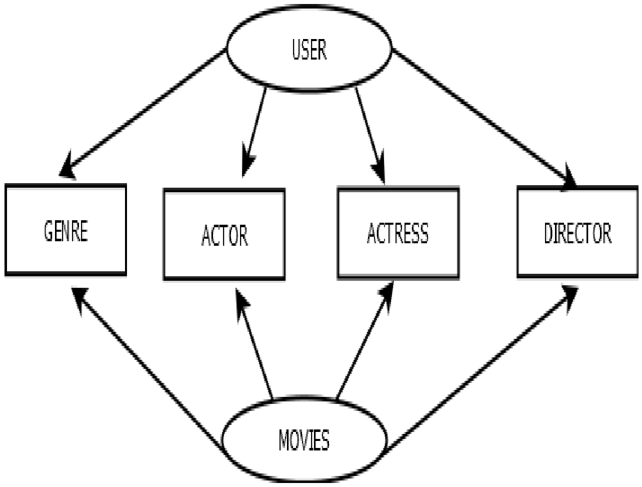

Figure 4 : Program Factors

The mathematical formulae used to calculate the positive and negative bias is given by Ψ, which represents the bias present in the data.

$$\Psi = 0.54\alpha + 0.46\beta \qquad \ldots(1)$$

Here, '*α*' refers to the Genre Score and '*β*' refers to the Actor/ Director Score.

The impact ratio of genre to director/actor score i.e. 54: 46 has again been inferred using the user survey sample conducted for over thousands Facebook® users. Given below is the step-by-step process for the implementation flow of the framework.

Algorithm 1: To calculate the Adjusted Sentiment Score.

i. *Input*

A corpus of movie review comments.

ii. *Variables*

An initially empty set of comments c,An initially empty set of tags t, which comprise of the three keyword types described above i.e. the keywords representing actor/actress, movie names and movie genre. An initially empty set of sentiments s.

iii. *Output*

Adjusted Sentiment Score *A*

```
$C$ ← retrieve_comments( $c_i$ )
For each $c_i$ □ $C$ do
    $s_i$ ← retrievesentiment($c_i$);
    $pos_i$ ← retrievepos($c_i$);
    $neg_i$ ← retrieveneg($c_i$);
    $A$ ← adjsentscore( $s_i$, $pos_i$, $neg_i$)
return '$A$'
```

The movie comments are collected in a set *c*. The index *i* refers to the fact that i^{th} comment is being processed. The total number of comments is taken to be *n*. For each comment in the set the keyword extraction is done and tags are collected in another set *t*. These tags are used to get the negative inclination score. Each comment is also manipulated to extract the sentiment types which are collected in a set *s*. The variables defining the positive bias score, negative bias score and sentiment are passed on to the adjusted score function to get the composite score.

Step 6: Unadjusted Score Calculation

The unadjusted score gives the subjective score of the user sentiments. This score needs to be refined to get an objective adjusted score.

The unadjusted score is calculated using ALCHEMY API framework that is incorporated in the movies reviews website. This score is calculated by applying the ALCHEMY API algorithm on the user comments in the website.

The unadjusted score thus calculated is given by *S*. Here, the number of users is taken to be m, while of multiple posts by a single user a variable n is used to keep a count of comments. The score of a single user is hence represented by,

$$S = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \ldots(2)$$

Step 7: Adjusted Score Calculation

To incorporated individual bias we look at three different possible aspects. Firstly, the positive inclination or the positive bias which shows overtly promoting behavior of the source. Secondly, the negative inclination or negative bias, which shows the detractor behavior of the source. Thirdly, when there is a mixed response of both positive and negative inclination by the source.

iv. *Alchemy Api*

The ALCHEMY API is then used to evaluate the unadjusted score for the user. The bias is incorporated in the score after the implementation of the positive and negative bias algorithms.

IV. Conclusion

The current systems lack the ability to objectively review a product based on user comments. This is because of the inherent biasness present in their comments. We combat this biasness problem by incorporating the positive and negative bias component in the existing unadjusted sentiment score of the object using various proposed algorithms. We calculated the degree of biasness an individual has towards or against the review parameters for the movie reviews corpus namely actor, director and genre. Finally, the system functions well in various scenarios like presence of only positive inclination of the user, presence of only negative inclination of the user, presence of both positive and negative inclination of the user and neutral attitude of the user. Hence, our system computes an objective score sans any individual bias.

V. Future Scope

The principal contribution of our research is the implementation of individual bias in the existing sentiment analysis algorithms. This can be used in various fields like business, journalism, product development among others. The research can be implemented across different algorithms and languages too.

The future endeavor in this direction would be implementation of unexplored biases in the system like selection bias, cognitive bias, first impression bias, herd bias, etc.

References Références Referencias

1. Kim, S., & Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the international conference on computational linguistics (COLING 2004) East Stroudsburg, PA,1367.
2. Cui, H., Mittal, V., & Datar, M. (2003). Comparative Experiments on Sentiment Classification for Online Product Reviews, 1265–1270.) S, S. K. (2013). Sentiment Analysis Based Approaches for Understanding User Context in Web Content Dept of Information Technology. doi:10.1109/CSNT. 2013.130
3. K. Dave, S. Lawrence, and D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web, pages 519–528. ACM, 2003
4. Wu, M., Wang, L., & Yi, L. (n.d.). A Novel Approach Based on Review Mining for Product Usability Analysis, (71), 942–945.
5. Gil De Zuniga, H., Puig-I-Abril, E., & Rojas, H. (2009). Weblogs, traditional sources online and political participation: An assessment of how the internet is changing the political environment. New Media and Society, 11, 553–574 Sciences, S., Lin, Y., Margolin, D., Keegan, B., & Lazer, D. (n.d.). Voices of Victory : A Computational Focus Group Framework for Tracking Opinion Shift in Real Time Categories and Subject Descriptors, 737–747.
6. Park, Ko, Kim, Liu, & Song, 2011) in Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications, 40(10), 4065–4074. doi:10.1016/j.eswa.2013.01.001
7. Larsson, A., & Moe, H. (2011). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media and Society, 14, 727–747.
8. Lin, E., Fang, S., & Wang, J. (2013). Mining Online Book Reviews for Sentimental Clustering. 2013 27th International Conference on Advanced Information Networking and Applications Workshops, 179–184. doi:10.1109/WAINA.2013.172
9. Wen, B., Dai, W., & Zhao, J. (2012). Sentence Sentimental Classification Based on Semantic Comprehension. 2012 Fifth International Symposium on Computational Intelligence and Design, 458–461. doi:10.1109/ISCID.2012.275
10. Yi-Fen Chen, Herd behavior in purchasing books online,, Computers in Human Behavior 24 (5) (2008) 1977–1992.
11. Guillaume Deffuant, Sylvie Huet, Collective increase of first impression bias,Complexity 15 (5) (2009) 25–3362
12. Piramuthu, S., Kapoor, G., Zhou, W., & Mauw, S. (2012). Input online review data and related bias in recommender systems. Decision Support Systems, 53(3), 418–424. doi:10.1016/j.dss.2012.02.006
13. Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of Natural Language Processing.Liu, Y., Huang,
14. Wu, M., Wang, L., & Yi, L. (n.d.). A Novel Approach Based on Review Mining for Product Usability Analysis, (71), 942–945.
15. Mishra, N., & Jha, C. K. (2013). Restricted Domain Opinion Mining in Compound Sentences. 2013 International Conference on Communication Systems and Network Technologies, 616–620. doi:10.1109/CSNT.2013.132
16. Bencz, A. (2012). Content-Based Trust and Bias Classification via, 41–47.
17. Cipriani, M., & Guarino, A. (2012). Estimating a Structural Model of Herd Behavior in Financial

Markets. SSRN Electronic Journal. doi:10.2139/ssrn.2080234

18. Sikora, R. T., & Chauhan, K. (2012). Estimating sequential bias in online reviews: A Kalman filtering approach. Knowledge-Based Systems, 27, 314–321. doi:10.1016/j.knosys.2011.10.011
19. Wang, C. A., Zhang, X. M., & Hann, I. (n.d.). Social Bias in Online Product Ratings : A Quasi-Experimental Analysis, (February 2010), 1–35.
20. Piramuthu, S., Kapoor, G., Zhou, W., & Mauw, S. (2012). Input online review data and related bias in recommender systems. Decision Support Systems, 53(3), 418–424. doi:10.1016/j.dss.2012.02.006
21. Jusoh, S., & Alfawareh, H. M. (2013). Applying fuzzy sets for opinion mining. 2013 International Conference on Computer Applications Technology (ICCAT), 1-5. doi:10.1109/ICCAT.2013.6521965
22. Lin, E., Fang, S., & Wang, J. (2013). Mining Online Book Reviews for Sentimental Clustering. 2013 27th International Conference on Advanced Information Networking and Applications Workshops, 179–184. doi:10.1109/WAINA.2013.172
23. Huang, Y., & Lin, H. (2013). Web product ranking using opinion mining. 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 184–190. doi:10.1109/CIDM.2013.6597235
24. Wu, M., Wang, L., & Yi, L. (n.d.). A Novel Approach Based on Review Mining for Product Usability Analysis, (71), 942–945.
25. Mishra, N., & Jha, C. K. (2013). Restricted Domain Opinion Mining in Compound Sentences. 201363