Soft Comput (2016) 20:3373–3379
DOI 10.1007/s00500-015-1779-1

FOCUS

Sentiment analysis of movie reviews: finding most important movie aspects using driving factors

Viraj Parkhe[1] · **Bhaskar Biswas**[1]

Published online: 18 July 2015

Abstract The opinion conveyed by the user towards the movie can be understood by sentiment analysis of the movie review. In the current work we focus on finding the aspects of a movie review which direct its polarity the most. This is achieved using certain driving factors, which are scores given to the various movie aspects. Generally its found that aspects with high driving factors affect the review polarity the most.

Keywords Sentiment analysis · Aspect-based sentiment analysis · Naive Bayes classifier · Aspect importance

1 Introduction and related works

Machine learning and/or various language processing tools are used to find whether the given document is positive or negative in polarity. This is called sentiment analysis. In the following work we have found whether a movie review is positively or negatively oriented using sentiment analysis. Document level sentiment analysis and aspect level sentiment analysis are the two levels of sentiment analysis (Singh et al. 2013; Parkhe and Biswas 2014). The first level uses certain lexicon-based method or machine learning approaches for document classification. Pang et al. (2002) suggested a good method of sentiment classification using Nave Bayes, SVM and Maximum Entropy classifiers. They experimented with different features like unigrams, unigrams and bigrams, adjectives, top unigrams, etc. and compared their results. Kang et al. (2012) proposed a method for mitigating the error caused when the accuracies of the positive and negative classes are expressed as average values. For this they proposed an improved Naive Bayes algorithm that reduced the accuracy gap. Sometimes the accuracy obtained by the machine learning algorithms is low; thus to address this problem Basari et al. (2012) used Support Vector Machines coupled with Particle Swarm Optimization to increase the overall accuracy. In their study they increased the accuracy from 71.87 to 77 %.

The second level deals with each individual aspect of the movie. A movie has many different aspects such as Direction, screenplay, acting, story, etc. and the reviewer may tend to give his/her opinion based on these aspects. Better analysis of the review is possible if individual aspect polarities are taken into consideration. Reviewers tend to have different opinion about various movie aspects. Thus for detailed analysis of the review, aspect-based analysis is the way to go. Many researchers have worked on aspect-based sentiment analysis. Thet et al. (2010) proposed a method for fine-grained analysis of sentiment orientation and sentiment strength of the reviewer towards the various aspects of the movie. It uses domain-specific and generic opinion lexicons to score the words and with the help of dependency tree, it identifies various inter word dependencies and helps in propagating the word score over the entire document. Singh et al. (2013) gave a new feature-based heuristic for aspect level sentiment analysis. In their scheme they analyse the review text and assign sentiment label on each aspect of the review. Then each aspect text is scored using SentiWordNet (2015) with feature selection comprising of adjective, adverbs, verbs and n-gram features. The overall document is then scored based on the aggregate score of each aspect. Yu et al. (2011) proposed a method for identifying important aspects from online con-

Communicated by S. Deb, T. Hanne and S. Fong.

✉ Viraj Parkhe
viraj.parkhe.cse10@itbhu.ac.in

[1] Department of Computer Science and Engineering, IIT-(BHU), Varanasi, India

sumer reviews. They identified the important aspects based on the observations that such aspects are commented the most in a review and overall product opinion is greatly influenced by consumer opinion on such important aspects. In their algorithm they formulate the aspect value distribution via a Multivariate Gaussian Distribution. In this paper, we tend to find the movie aspects that dictate the polarity of the review the most. For this we give different weightage to individual movie aspects, called driving factors. The overall score is the sum of individual aspect scores weighed by their driving factors. The approach of Yu et al. (2011) differs from our approach in the method by which they assign aspect values. They use a Multivariate Gaussian distribution while we use a randomized approach to assign values to the driving factors. Also we choose those driving factors that give the maximum accuracy as the best driving factors. The rest of the paper is organized as follows: Sect. 2 describes the proposed method; Sect. 3 gives the dataset, experimental results and performance; Sect. 4 gives the conclusion and future work, and the last section gives Compliance with Ethical Standards and references.

2 Proposed method

The following method suggests a technique for aspect-based sentiment analysis of movie reviews (Parkhe and Biswas 2014). Figure 1 describes the method flow. The first step is pre-processing. In this step, we collect reviews from different sources and pre-process them to make them suitable for use in the method. Pre-processing includes formatting of reviews so that they can be aligned in the required format. For this the HTML tags and other tags were removed. For the following method the reviews were pre-processed into simple text format. The next step was to separate the review text into aspects and this was done using Aspect Based Text Separator (ABTS). The various movie aspects that we used are screenplay, music, acting, plot, movie and direction. An aspect-specific lexicon was used to separate the review aspect wise. Table 1 shows some of the words used to separate the sentences (Thet et al. 2010). Each word in the lexicon was associated with the part of speech of that word. While searching the sentence to match the lexicon word, we first tagged the sentence with the Stanford Part of-Speech Tagger (2015) and then we matched the lexicon word within the sentence having the same part of speech.

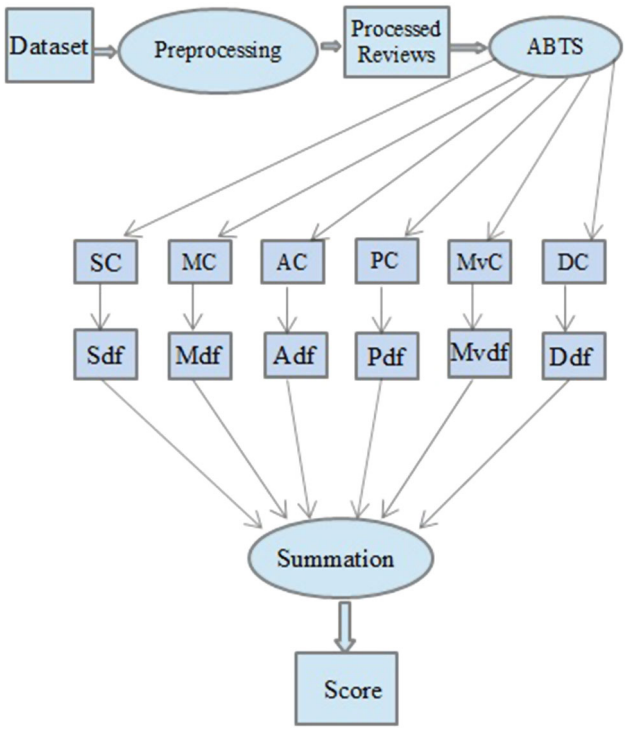

Fig. 1 Diagram for the proposed method

Table 1 Lexicon used for aspect based text seperator

Aspect	Aspect words
Screenplay	Scene, scenery, animation, violence screenplay, action, etc
Music	Music, score, lyric, sound, audio, musical title track, etc
Acting	Acting, role playing, act, actress, actor role, portray, character, villain performance, etc
Plot	Plot, story, storyline, tale, romance dialog, script, storyteller ending, storytelling revenge,betrayal,writing, etc
Movie	Movie, film, picture, moving picture motion picture, show, picture show, pic flick, romantic comedy, etc
Direction	Directing, direct, direction, director, filmed filming, film making, filmmaker, cinematic edition, cinematography, etc

In the next step, these separated reviews were forwarded to the aspect-specific classifiers. A Naive Bayes classifier Pang et al. (2002) was used for this purpose. It calculates the probability of a word or albeit a sentence, belonging to positive or a negative class of reviews. The outputs were obtained using the traditional training and testing method. The outputs were either -1 or 1 denoting that the input text was negatively or positively oriented, respectively. Instead of NB we can use any classifier like SVM, etc. that is able to clearly classify the text into two classes. However, we must carry out proper processing of input data so that it meets the proper data format requirement for each classifier. Based on the weightage of the driving factors of the movie, the aspect-based output is multiplied with the respective driving factor.

The higher the value of the driving factor of an aspect, the more is its importance in the review. The driving factors follow the relationship

Table 2 Performance measures

Sr. no	Accuracy	Recall	Specificity	Precision
1	0.79372	0.76568	0.82176	0.81117
2	0.78956	0.75888	0.82024	0.80848
3	0.78268	0.70512	0.86024	0.8345
4	0.77996	0.75176	0.80816	0.79669
5	0.76912	0.75192	0.78362	0.7787
6	0.7598	0.73184	0.78776	0.7751
7	0.74692	0.72312	0.77072	0.75926
8	0.7358	0.7156	0.756	0.74572
9	0.7254	0.6872	0.76368	0.74410
10	0.71812	0.68672	0.74952	0.73273

$$\sum \alpha_i = 1,$$

where $\sum \alpha_i$ is the (ith) driving factor. The net output obtained is the sum of all the classifier outputs obtained multiplied with their respective driving factors. The output is

$$\omega(\mathrm{d}) = \sum \alpha_i X_i \quad X_i \subseteq [-1, 1],$$

where α_i is the driving factor of (ith) aspect and X^i is the output of (ith) classifier and (d) is the document under consideration. Now if

$$\omega(\mathrm{d}) \leq 0 \rightarrow \text{negative classification of review} \quad \mathrm{d}$$
$$\omega(\mathrm{d}) > 0 \rightarrow \text{positive classification of review} \quad \mathrm{d}$$

Thus we have used a threshold score for the classification of the document.

3 Dataset, experimental results and performance

The dataset was acquired from the Large movie review dataset site of Stanford AI Lab (Maas et al. 2011; Large Movie Review Dataset 2015; Parkhe and Biswas 2014). The dataset consists of 25,000 positive and 25,000 negative reviews and was collected from IMDB. Though there is no specific time span for review collection from IMDB, but it was ensured that no more than 30 reviews from a single movie get included in the final dataset. Because of even number of positive and negative reviews, the minimum accuracy that we can obtain from the experiment is 50 %. The dataset contains only highly positive and highly negative reviews. The authors of the dataset included a negative review only if it scored 4 out of 10 and included a positive review if it scored 7 out of 10 on a benchmark set by them (Maas et al. 2011). Neutral reviews were omitted. It was seen that ABTS separated the review into various aspects having unequal text distribution. This was due to the fact that in each review, the reviewer commented on each aspect in unequal number of sentences. Also in some reviews not all aspects were commented on. The score for such reviews was made 0. As mentioned in the previous section a Nave Bayes classifier was used for classifying the separated aspect based text. The individual classifiers got the aspect-based text as input in the ratio of 70:30, for training and testing, respectively. The experiment ran for 1000 iterations and during each iteration, random values between 0 and 1 were assigned to the driving factors. For the particular dataset under consideration, the driving factors giving the highest accuracy were chosen as the best driving factors (Table 2). The experiment conducted gave results as depicted in Table 3.

The results in Table 3 depict the relationship between accuracy and driving factors used. The highest accuracy obtained was 0.79372, i.e. 79.372 %. The corresponding factors are Screenplay—0.07877, Music—0.11756, Acting—0.28147, Plot—0.16390, Movie—0.31225 and Direction—0.108133.

Thus by using the mentioned driving factors, we get an accuracy of 79.372 %. This is the highest accuracy obtained using this method. Also its worth noting that giving equal importance to all factors, i.e. giving each a value of 0.165,

Table 3 Performance measures (DF = driving factor)

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.79372	0.07877495	0.1175615	0.218479825	0.16390358	0.31225607	0.10813343
0.78956	0.047883925	0.004667212	0.2291204	0.20019746	0.30320117	0.2136368
0.78268	0.165	0.165	0.165	0.165	0.165	0.165
0.77996	0.004460782	0.013207557	0.43918112	0.17664995	0.3274761	0.038970188
0.76912	0.07476745	0.5480689	0.09583495	0.1124063	0.1001353	0.02502478
0.7598	4.44E−05	0.001835718	0.4903199	8.41E−05	0.042956525	0.463557
0.74692	0.002677993	1.86E−04	0.9551522	0.025326777	0.010514759	0.006046253
0.7358	0.001611796	0.49006925	0.01817123	0.003027879	0.010863352	0.473160775
0.7254	0.003411332	0.000922476	0.001115599	0.49038495	3.75E−04	0.5032535
0.71812	0.24075805	0.198029235	0.00240013	0.003753847	2.46E−04	0.55194675

has resulted in a lower accuracy of 78.268 % than the highest accuracy obtained by unequal distribution of factors. The effect of changing driving factors can be seen in the accuracy of the overall classification obtained. In the above case of 79.372 % accuracy we have given most importance to the Movie, Acting and Plot aspects. Thus we can interpret from the results that in the reviews used from the dataset, the user has given more importance to these factors while writing the review. It also means that if the reviewer gives a positive opinion towards these aspects, then due to their high importance the overall review will tend to be positive even if he/she gives a negative opinion towards the other aspects. Giving more importance to certain factors also has an added advantage; it tends to suppress the user opinion about other factors. Suppose we have a review X and it contains user opinion about two factors F1 and F2. Also the overall orientation of the review is positive in nature. The user has given a positive review about F1 and a negative about F2. Also the amount of text in the review for F1 aspect is less as compared to the F2 aspect. Now if we use any non-aspect based sentiment analysis method, then since text size of F2 is greater than text size of F1 and also since F2 is negative in orientation, the overall document score will tend to reduce and skew towards negativity. On the other hand, if driving factors are used and F1 is given more importance the document score will better reflect the positivity of the review. Since each aspect of a movie is analysed separately in this method, we can track the effect each aspect has towards the overall score of the document. This individual aspect-based tracking can be used in a fine-grained aspect-based recommendation system, which recommends movies based on their various aspects instead of the overall rating of the movie. Also this method can be applied on a product review dataset thus enabling us to see what opinion each user has on the various aspects of the product, thus helping in the development of proper product placement strategy. It is very difficult to acquire such in-depth knowledge from the dataset using non-aspect based methods.

We wanted to see how the above method would work on reviews of specific movie genre. Thus we applied the method on movie reviews of genres like action, adventure, animation, comedy, crime, documentary, drama, horror and the results obtained are showed in Tables 4, 5, 6, 7, 8, 9, 10. For certain experimental simplifications, the sum of the driving factors is taken to be 2 instead of 1 as mentioned previously. As can be seen from the tables, we got an accuracy of 63.8 % for action genre, 63.33 % in Adventure genre, 81.48 % in animation genre, 77 % in Comedy genre, 87.3 % in Crime genre, 84.82 % in Documentary genre, 76.64 % in Drama genre and 83.33 % in Horror genre.

Table 4 Experimental results for action genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.62857145	0.2607335	0.09698348	0.080310054	0.48212352	0.106518164	0.86561745
0.63809526	1.4502046E−5	0.01205827	1.7990339	0.023204818	1.1991088E−4	0.16554123
0.5952381	0.030099688	0.62651104	0.4437882	0.19853848	0.6017535	0.06541103
0.60952383	0.017110074	0.0276303	0.6291681	0.185065	0.015373883	1.0696439
0.6142857	0.0050853747	0.0287022	0.0068396833	1.3842763	0.2746869	0.27360862

Table 5 Experimental results for adventure genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.54444444	0.16695978	0.019693026	0.2845779	0.117406845	0.044530526	1.2983667
0.600000	0.0062500793	0.43146822	0.37962252	1.0956551	0.005007505	0.0631768
0.500000	0.101619005	0.068823755	1.7743889	1.2859914E−5	0.054626025	4.665379E−4
0.6111111	0.018982153	0.023985077	0.30932263	1.2149528	0.110125825	0.32231408
0.6333333	1.0164111	0.20472622	0.16184342	0.36442122	0.01203593	0.23595665

Table 6 Experimental results for animation genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.762963	0.40278578	0.13121434	0.1639517	0.16147086	0.023833089	1.0766937
0.7259259	0.015430119	0.01450169	0.0018918073	1.8796889	0.06999449	0.0032109926
0.8148148	1.9827418	0.0044073765	0.002998911	0.004753623	0.0013958901	0.002580362
0.742163	0.049183633	0.026727557	0.15083629	0.6253178	0.014134193	1.1253048
0.78763	0.0024441832	0.001321153	0.47855914	1.4879216	5.6015153E−4	0.019595714

Table 7 Experimental results for comedy genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.6869806	0.029056318	0.08829725	0.03141967	0.018376663	0.034200396	1.7722781
0.74515235	0.98266107	0.10399312	0.19353703	0.6441466	0.0032735833	0.06635751
0.72853184	5.1874836E−4	1.2089423	0.77944654	0.0042970385	7.162274E−5	0.006487943
0.7617729	0.0031588415	0.3166019	0.23175706	0.29527476	0.27282205	0.82218164
0.7700831	0.43993464	0.7786206	0.0026597457	0.2658547	0.50571316	0.004942937

Table 8 Experimental results for crime genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.82539684	0.046272777	0.0028346716	0.0023228936	0.043214377	1.8829044	0.009109251
0.8730159	0.23899242	0.39356804	0.24650577	0.116251566	0.8519834	0.06261088
0.85714287	5.160264E−5	0.81093293	0.0015740955	1.1405741	0.045843605	3.704426E−4
0.7936508	0.09719603	6.024262E−4	1.899839	0.0010096797	2.9936084E−4	8.913378E−4
0.84126985	0.8175193	0.044794194	0.001806062	0.1312579	0.9157112	0.07478792

Table 9 Experimental results for documentary genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.84285713	0.60168445	0.7291384	0.0066952403	0.09652837	0.12207278	0.39357853
0.83035713	0.03726328	0.008822674	0.027112441	0.2241178	0.33716756	1.2007562
0.84821427	0.40020192	1.1154401	0.038128346	0.07422164	0.0016440379	0.3168806
0.78571427	0.013991571	1.7053754	0.013337747	0.025600998	0.08954371	0.10341096
0.79464287	0.013194267	0.01776664	0.38369128	1.4746889	0.036601648	0.047439173

Table 10 Experimental results for drama genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.7639594	0.08935903	0.21936458	0.13797145	0.05258798	1.3640177	0.10569177
0.75888324	0.17959706	0.04889245	1.5469579	0.09408904	0.026836606	0.05110698
0.76649743	0.06287201	0.0021581356	0.016743837	0.032732528	1.698626	0.18364914
0.74873096	1.5242645	0.0022528782	0.06901909	0.29797488	0.07420794	0.017557843
0.7614213	0.20333073	1.025873	0.21291313	0.003925945	0.5055761	0.045527093

The various performance measures used were (Singh et al. 2013)

$$\text{Accuracy} = \frac{\text{Total correctly classified documents}}{\text{Total number of documents}}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Specificity} = \frac{tn}{\text{Total negatively oriented documents}}$$

$$\text{Recall} = \frac{tp}{\text{Total positively oriented documents}},$$

where (tp), (fp) and (tn) are the true positives, false positives and true negatives obtained during the classification. The result obtained by applying the various performance measures can be seen in the given tables. As can be seen from Fig. 2 for action genre we got direction, plot and screenplay as the most important driving factors, for adventure genre we got direction, acting and screenplay, for animation genre we got direction, screenplay and acting, for comedy we got direction, music and movie, for crime we got movie, screenplay and plot, for documentary we got music, screenplay and direction, for drama we got movie, music and acting and for horror we got acting, movie and direction as the most important driving factors. Only the highest accuracy across each genre was considered for obtaining the above results. The graph denotes the percentage distribution of the driving factors across each genre. The total value of these factors comes out to be 2 as stated previously. Thus it shows that each genre has unique important driving factors and if the reviewer comments positively on these aspects then the overall accuracy of the classification increases (Table 11).

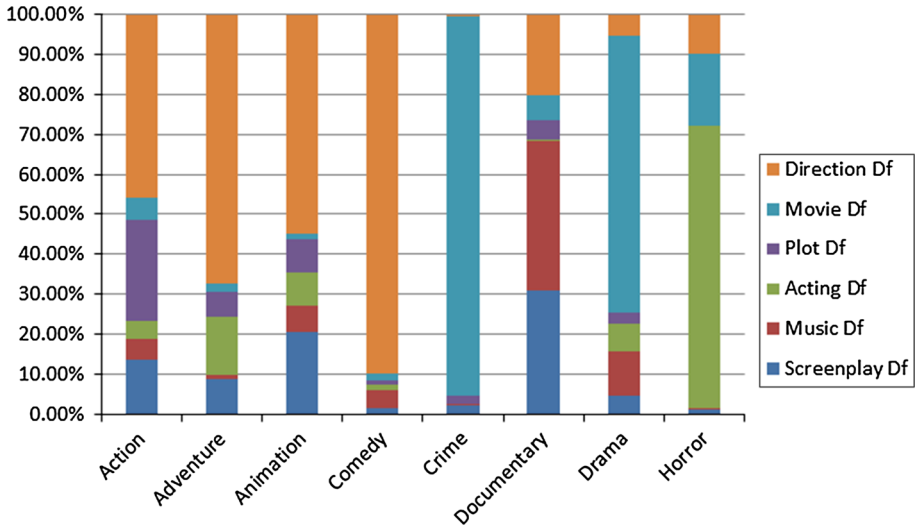

Fig. 2 *Bar* graph showing the distribution of driving factors across all genres

Table 11 Experimental results for horror genre

Accuracy	Screenplay (DF)	Music (DF)	Acting (DF)	Plot (DF)	Movie (DF)	Direction (DF)
0.7619048	0.019708144	0.008453023	1.4132087	0.0027013326	0.3546674	0.19764264
0.8333333	0.017502377	0.48868188	0.59327596	0.13216278	0.72132105	0.042207204
0.6666667	0.005021711	0.019522838	0.6477796	0.75494176	0.013859565	0.5446752
0.7380952	0.0017782062	0.24107058	1.7099496	0.024216007	0.0023502277	0.0018495012
0.61904764	1.4361262	0.31911767	0.02072717	0.18048742	0.015468956	0.014186056

4 Conclusion and future work

The experiment was conducted to find which movie aspects influence the orientation of the review using driving factors. It concluded with Movie, Acting and Plot aspects getting overall high driving factors and resulting in an accuracy of 79.372 % for the current dataset in consideration. The importance of these aspects may or may not change, but since the experiments were conducted on a large dataset, it is quite unlikely that it will.

As we can see from the results obtained for genre-specific reviews, the method gave high accuracy for some genres, while it gave lower accuracy for others. Thus its evident that the method used for mixed review classification is not that good for reviews of certain genres. Thus a newer approach need to be developed of genre-specific classification of reviews as reviews of different genres tend to incorporate genre-specific words or sentences that can have different meaning based on the context in which they are used. For instance, the word funny is used in a good context for a comedy movie but may be used in a wrong context for movie genre like horror, etc. Thus such context-specific words and sentences resulted in uneven accuracy as depicted in the results.

The current method used for classifying the text is Naive Bayes Classifier which uses a bag-of-word approach. This approach does not consider the inter word meaning dependencies and also the context in which the word was used, i.e. genre. For this purpose we tend to develop scoring method using context specific lexicon. Each word in the lexicon will have a different positive and negative score based on the context (genre) in which it was used. Also to incorporate the inter word dependencies we tend to use clause-based scoring of a sentence. It scores each clause of a sentence individually and thus the overall sentence score is the sum of individual clause scores. Thus by coupling the above improved method with the use of genre-specific driving factors we tend to obtain more refined scores for the movie reviews.

Acknowledgments The above work is an extension of previous work published in ISCMI 2014 (Parkhe and Biswas 2014). Proper citations have been included for the same in the above work for transparency purposes.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest

Informed consent Also Informed consent was obtained from all individual participants included in the study. This article does not contain any studies with human participants or animals performed by any of the authors.

References

Basaria ASH, Hussina B, Anantaa IGP, Zeniarjab J (2012) Opinion mining of movie review using hybrid method of support vector

machine and particle swarm optimization. In: Malaysian Technical Universities conference on engineering & technology (MUCET 2012) part 4: information and communication technology

Bro J, Ehrig H (2010) Generating a context-aware sentiment lexicon for aspect-based product review mining. In: IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology

Kang H, Yoo SJ, Han D (2012) Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst Appl 39(5):6000–6010

Large Movie Review Dataset (2015) Acquired from stanford AI lab. http://ai.stanford.edu/~amaas/data/sentiment

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learningtechniques. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Philadelphia, pp 79–86

Parkhe V, Biswas B (2014) Aspect based sentiment analysis of movie reviews: finding the polarity directing aspects. In: Proceedings of international conference on soft computing and machine intelligence 2014

Sentiment Analysis (2015) Wikipedia-the free encyclopedia. http://wikipedia.org/wiki/Sentiment_analysis

SentiWordNet (2015) Lexical resource for opinion mining. http://sentiwordnet.isti.cnr.it

Singh VK, Piryani R, Uddin A, Waila P (2013) Sentiment analysis of movie reviews a new feature-based heuristic for aspect-level sentiment classification. In: Proceedings of the 2013 international muli-conference on automation, communication, computing, control and compressed sensing, Kerala-India, IEEE Xplore, pp 712–717

Stanford Part-Of-Speech Tagger (2015) Stanford natural language processing group. http://nlp.stanford.edu/software/tagger.shtml

Taboada M, Brook J, Stede M (2009) Genre-based paragraph classification for sentiment analysis. In: Proceedings of SIGDIAL 2009: the 10th annual meeting of the special interest group in discourse and dialogue, University of London, Queen Mary, pp 62–70

Thet TT, Na J-C, Khoo CSG (2010) Aspect-based sentiment analysis of movie reviews on discussion boards. J Inf Sci 36(6):823–848

Yu J, Zha Z-J, Wang M, Chua T-S (2011) Aspect ranking: identifying important product aspects from online consumer reviews. In: Proceedings of the 49th annual meeting of the association for computational linguistics, Portland, Oregon, pp 1496–1505, 19–24 June 2011