Research Journal of Applied Sciences, Engineering and Technology 10(6): 623-629, 2015
DOI:10.19026/rjaset.10.2471
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: January 1, 2015 Accepted: February 11, 2015 Published: June 20, 2015

Research Article

Review Content Analytics for the Prediction of Learner's Feedback with Multivariate Regression Model

T. Chellatamilan, B. Magesh and K. Balaji
Department of Computer Science and Engineering, Arunai Engineering College, Tiruvannamalai 606603, Tamilnadu, India

Abstract: E-learning facilitates both synchronous and asynchronous learning and it plays very important role in the teaching learning process. A large group of learners are engaged in the idea exchange independently by interacting with the members present in the learning management system. In order to generate meaningful learning outcome of the individual peer learners, the feedback review is very essential to extract the conceptual content which reflect the instantaneous learner's behavior, emotions, capabilities, interestingness and difficulties and to fits them effectively. Collecting feedback in the form of numeric scale is very tough for both the learners and facilitators while specifying the rating, but it is too easy for the learners provide feedback in the form of text messages. The key challenge for analyzers is to extract the meaningful feedback content and dynamic rating of the learner's feedback related to various conceptual contexts. We propose a novel method using multivariate predictive model for conceptual content analytics based on e-learners reviews using standard statistical model inverse regression. Finally the analysis is used in the prediction studies and to illustrate their effectiveness against the learner's feedback.

Keywords: E-learning, feedback, logistic regression, review analytics, text mining

INTRODUCTION

In this study, we study and propose a new text analytics problem named "MPMCA" (Multivariate Predictive Model for Conceptual Content) which is aiming at interpreting textual data expressed about the various issues in an online e-learning review discussion forum at the level topical analysis to discover the point scale of each categorical review feedback. Generally the feedback of the learners in an e-learning system deals about the textual interpretation of the feedback. It is so easy for the reviewer to give the review feedback in terms of text, but it is too hard for the assessor of the review to interpret the ratings of such textual review. This system automatically assigns the rating score for the feedback according to the textual lexicon words used in the respective topic of review feedback. We propose a modern statistical and probabilistic model to analyze the textual content and then to predict the best point scale (numeric value) represents the descriptive feedback through inherent regression model.

This model can also supports similar systems where the reviews are collected in the form of text such as entity rating, opinion rating, restaurant reviews, on line shopping and etc. In the emerging advancement of semantic web, more and more users of internet shows their interest, opinion, sentiment on all kind of services in the internet of things. The Service provider may use this review content to improvise the quality of services. Only the challenging issue on this kind of system is the interpretation and the assessment of such point scale value. More over the volume and the format of such reviews grew very rapidly so that the review analytical process becomes difficult for review analyzer. Consider a typical student review collection format shown in the Table 1. This review deals with multiple topics of teaching learning process such as teachers information, learning material information, question paper details and etc., the students only submits the rating of each topical category of the reviews in terms of both in textual form as well as point scale rating. This has been consider as the training set for training the prediction model and in turn this prediction model automatically estimates the point scale score of all the reviews of the test set collections. To achieve such detailed understanding of feedback, we propose a text analytical method named "MPMCA". With the help of all such set of training feedback, the MPMCA aims at predicting the point scale value of each feedback review at the level of individual topical category to estimate each learners inherent rating. The point scale weights are directly useful for analyzing the difficulties of all learners.

In this study, we study and propose a new text analytics problem named "MPMCA" (Multivariate Predictive Model for Conceptual Content) which is

Corresponding Author: T. Chellatamilan, Department of Computer Science and Engineering, Arunai Engineering College, Tiruvannamalai 606603, Tamilnadu, India

Table 1: The general architecture of MPMCA

Topical aspects	Summary	Ratings (5 point scale)
Student • Language problem • Availability of contents • Illness(self/family) • External events	I have communication problem while learning the topic. Excellent learning contents are given by the faculties. Due to the viral fever I am unable to prepare well for the examinations.	3.2
Teacher • Syllabus not covered • Problems not worked out • Poor vocabulary/audible	Though the syllabus is vast the teacher have covered entire chapters and solved more problems in the class with sound communication.	4.0
Question paper • Tough • Out of syllabus	The question paper is very easy and is compared with earlier question set. The last unit question is unexpected.	3.5
Valuation • Tough • Revaluation applied	The valuation seems that it is easy because most of my friends have cleared the semester subjects. I have applied photo copy of my answer script.	2.7

aiming at interpreting textual data expressed about the various issues in an online e-learning review discussion forum at the level topical analysis to discover the point scale of each categorical review feedback. Generally the feedback of the learners in an e-learning system deals about the textual interpretation of the feedback. It is so easy for the reviewer to give the review feedback in terms of text, but it is too hard for the assessor of the review to interpret the ratings of such textual review. This system automatically assigns the rating score for the feedback according to the textual lexicon words used in the respective topic of review feedback. We propose a modern statistical and probabilistic model to analyze the textual content and then to predict the best point scale (numeric value) represents the descriptive feedback through inherent regression model.

This model can also supports similar systems where the reviews are collected in the form of text such as entity rating, opinion rating, restaurant reviews, on line shopping and etc. In the emerging advancement of semantic web, more and more users of internet shows their interest, opinion, sentiment on all kind of services in the internet of things. The Service provider may use this review content to improvise the quality of services. Only the challenging issue on this kind of system is the interpretation and the assessment of such point scale value. More over the volume and the format of such reviews grew very rapidly so that the review analytical process becomes difficult for review analyzer. Consider a typical student review collection format shown in the Table 1. This review deals with multiple topics of teaching learning process such as teacher's information, learning material information, question paper details and etc. The students only submits the rating of each topical category of the reviews in terms of both in textual form as well as point scale rating. This has been consider as the training set for training the prediction model and in turn this prediction model automatically estimates the point scale score of all the reviews of the test set collections. To achieve such detailed understanding of feedback, we propose a text analytical method named "MPMCA". With the help of all such set of training feedback, the MPMCA aims at predicting the point scale value of each feedback review at the level of individual topical category to estimate each learners inherent rating. The point scale weights are directly useful for analyzing the difficulties of all learners.

THE PROBLEM DEFINITION AND LITERATURE REVIEW

In this section, we formally define the problem of review text analytics. As an analytical problem, MPMCA assumes that the input is a set of feedback reviews of the interesting topical category of the student review. The Training review set has both the textual review and its corresponding point scale rating. Formally Let $R = \{r_1, r_2, r_3, ... r_{|R|}\}$ be the set of text review documents with multiple topical category and each review document r€R is associated with overall score Sr with n unique vocabulary set of the entire review $V = \{w_1, w_2, w_3, ... w_n\}$.we further consider that we are given topics which are rating factors that potentially affect the overall score of the given topical category. The topical score is specified through a few feature key words and provides the basis for the inherent topic score analysis. Informally MPMCA is aiming at discovery of the hidden latent topic rating and its weights.

The process of collecting feedback on their experience is widely recognized as a central strategy for monitoring the quality and standards of teaching and learning in Higher Education Institutions which set out the information about quality and standards of learning and teaching (Jara and Mellar, 2010). Assessment and feedback lies at the heart of the learning experience and forms a significant part of both academic and administrative workload. It remains however the single biggest source of student dissatisfaction with the higher education experience (Ferrell, 2012). The rating inferences solution evaluates the multi point scale representation of the interest based on multi class text categorization using metric labeling formulation

because the categorization is harder than ranking and vice versa (Pang and Lee, 2005).

The reading difficulty of predicting the rating or polarity of the descriptive feedback is always the problem for the persons who are giving or assessing the feedbacks regarding the particular subjects. The grade level of the phrase or paragraph can be estimated using the mixture of language models with the help of relatively labeled data (Callan, 2004). The semantic orientation of the phrases also good associations in classifying the reviews using unsupervised learning and Point Scale Mutual Information between two words where as the review contains the adjectives and adverbs (Turney, 2002). The sentiment of the review can also be assessed using the review argumentation among the discussions. From such textual arguments, the sentiment flow pattern can be structured and then the similarity of such pattern is compared with the peer review to estimate the score of such peer review (Wachsmuth *et al*., 2014).

The problem of automatic polarity mining refers to identifying and extracting topical information from natural language processing. The projection of n-gram into low dimensional latent semantic space devises a new embedding mechanism for sentiment review analysis (Bespalov *et al*., 2011). The portion of rate aspect text plays very important role in building model that best suit the prediction system. Since a review consists of multiple aspects, it becomes conflict for separating the reviews into individual aspects. The model is required to learn sentiment neutral lexicons of words which describe each of the aspects (McAuley *et al*., 2012). The inverse regression model discovers and quantifies the variations in topic expressions which influence the context on the relative prevalence of different topics in the document (Rabinovich and Blei, 2014). The multinomial inverse regression is introduced as an Information retrieval procedure for predictor sets that can be represented as draws from a multinomial and details its application to text-sentiment analysis. The generic regression does nothing to leverage the particulars of text data, independent analysis of many contingency tables leads to multiple-testing issues and pre-defined word lists are subjective and unreliable (Taddy, 2013).

The multinomial logistic regression model becomes specifically attractive leading to a monotonically converging sequence of iterations (Dankmar, 1992). The recommendation is performed by extracting the concepts during the conversations of the users and their queries with the help of semantic web technologies. It is able to automatically analyze the conversation chunks, identify treated topics and suggest all available material related to these topics and useful to enrich and get meaningful the conversation itself (Granito *et al*., 2014).

The multivariable frequency response analysis is focusing on the singular value decomposition; sensitivity functions, relative gain analysis and the role of multivariable right-half plane zero (Skogestad and Postlethwaite, 2007). The summarization technique involves different text inputs and which mostly

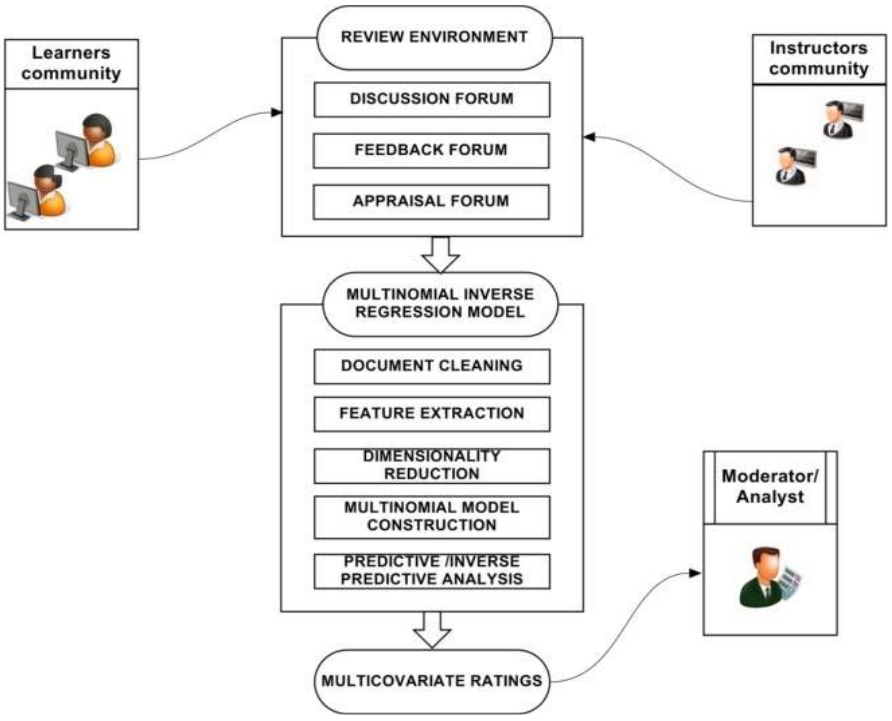

Fig. 1: The general architecture of MPMCA

considers only the features in the opinions of the products (positive and negative) but it fails in selecting the subset of the reviews to be considered to capture the key points in the text opinion and emotional summarization (Hu and Liu, 2004). The terms in topics are modeled by multinomial distribution; and the observations for a random field are modeled by Gibbs distribution (Sun *et al*., 2012). While this is feasible and gives users control over the aspects to be analyzed, there may also be situations where such keywords are not available (Wang *et al*., 2010). Document-level covariates enter the model through a simple generalized linear model framework in the prior distributions controlling either topical prevalence or topical content (Roberts *et al*., 2013). The review analytics is also analogous to the segmentation of blogs on the basis of topic labels provided by users, or topic discovery on the basis of tags given by users on social bookmarking sites (Titov and McDonald, 2008). Recently, the topic modeling can be used to improve the efficacy of baseline performance with multi aspect prediction of rating using Latent Dirichilet Allocation (LDA). The Overall polarity rating of such feedback review depends on all such individual topical aspects (Lu *et al*., 2011).

General architecture: The general architecture of the prediction model consists of fivefold subsystem as shown in the Fig. 1:

- Learners community
- Instructors community
- Moderator (Assesor)
- Multi nomial inverse regression model
- Multi-variate rating analysis

The Learners community and instructors community always shows their interestingness, feedback, difficulties or in convenience through the review environment like discussion forum, feedback forum and appraisal forum in the form of descriptive text with most predominant words for each polarity. The overall input for the prediction model is the collection of such review text. The second important sub system of this model is the Inverse Multinomial regression model. The Data pre processing is applied initially to make them ready for constructing the bag of words dataset along with the frequency count of each token after removing the stop words and stemming. Then the Bag of word dataset is applied to the topical model (LDA) to find out the latent topics or aspects discussed in the review text. The overall rating of the individual aspects or topics is calculated according to the number of keywords matched with the topics/aspects. The cumulative score is obtained by adding up all such score of each of the topics.

METHODOLOGY

There are many critical problems that can affect the learners academic experience and success such as language problem, external events, illness (self/family) and textbooks not available, poor in presentation and lack of preparation. The statistical analysis of word counts from high dimensional textual documents is the state of art. The lexiconization or tokenization of generating bag of words is the very first process of text analytics which assign frequencies to words or its combinations with other words. The stemming and stop word removal has been applied next for removing the morphological words includes the stop words. One of the most challenging issues in solving the problem of MPMCA is that we do not have detailed information about the hidden rating score on each of the topic through important keywords. In order to provide these challenges, we propose a modern statistical method using logistic regression model. The Topical segmentation is performed for a review document based on the keywords describing that topic. The online learning portal facilitates to collect various learners feedback summary as described in the Table 1.

Topic segmentation: The first step with topical segmentation is to map the review sentence of a review into sub sentences or subsets corresponding to each aspect.

Algorithm (topic segmentation):

Input: Review Collection Set R = {r1, r2, r3.... r|R|}
Topical Keywords {t1, t2, t3... tk}
Vocabulary V = {w1, w2, w3.... wn}.
Output: Review Split up into sententences with topical alignment.
Step 1: Divide all reviews into sentences S = {s1, s2, s3.... sm}.
Step 2: Match the topical keywords in each sentence S and record the matching topics with its count.
Step 3: Assign label to each topic and its maximum count.
Step 4: Calculate weight of each feature keyword in the vocabulary.
Step 5: Sort the words according to its count.
Step 6: Output the sorted list.

TF-IDF: The Term frequency and Inverse Document Frequency is a score base of screening of words in the document corpus as mentioned in Eq. (1):

$$TF - IDF(w,d) = f(w,d) * \log\left(\frac{n}{Dd}\right) \quad (1)$$

where,
f (w, d) = Frequency count of word ‘w’ in the document ‘d’

n = Total number of documents in the corpus
D_d = Number of documents containing the word 'w'

Bigrams: Bigrams are group of two adjacent neighbor words that are commonly occurs during the statistical analysis of text.

Multivariate logistic regression: Multinomial logistic regression is used to perform the analysis over the relationships between a non-measurable dependent parameter and measurable independent features. Multinomial logistic regression provides a collection of coefficients with all zero values for the reference group, matches to the coefficients for the reference set of a pseudo-coded parameter. Such equations can be used to measure the probability that subject is a element of each of the three sets. The predicted case is belonging to the group associated with the largest possibility. A case is predicted to belong to the group associated with the highest probability. The maximization and likelihood reduction model was used to test the relationship among the independent parameters of each of the sets.

Note that you can think of logistic regression in terms of transforming the dependent variable so that it fits an s- shaped curve. The odds ratio is the probability that a case will be a 1 divided by the probability that it will not be a 1. The logit is the natural log of odds and it is a linear function of the x's (that is, of the right hand side of the model).

The probability of dependent variable y for the given independent variable x is calculated as per the following Eq. (2):

$$P(y \mid x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad (2)$$

The regression coefficient is represented as β and the α represents the regression constants in Eq. (3):

$$\alpha+\beta x = \log\frac{P(y \mid x)}{1-P(y \mid x)} \quad (3)$$

The score is obtained as per the Eq. (4) and (5) from a set of weights that are linearly combined with the explanatory variables (features) of a given observation using a dot product: where X_i is the vector of explanatory variables describing observation i, β_k is a vector of weights (or regression coefficients) corresponding to outcome k and score (X_i, k) is the score associated with assigning observation i to category k:

$$Score(X_i, k) = P_k.X_i \quad (4)$$

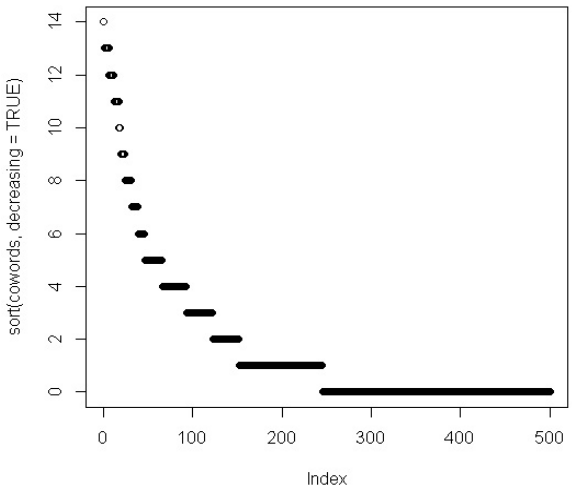

Fig. 2: Plot of words with its co-words

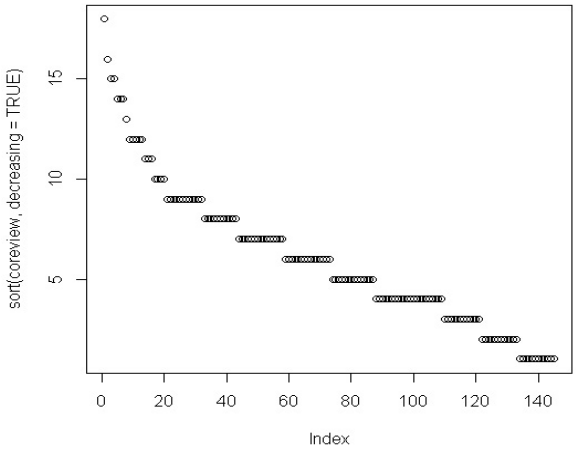

Fig. 3: Plot of words with its co-words in the overall review

The multinomial logistic regression uses a linear predictor function as expressed in the following Eq. (5):

$$f(k,i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \ldots\ldots\beta_{m,k}x_{m,i} \quad (5)$$

EXPERIMENTAL RESULTS

We have taken the movie lens dataset (train.tsv downloadable) for our experiments, contains 131000 reviews with counts on 25400 bigrams. It covers almost all the topical aspects of movie related genre information. We applied 70% of rows for training and the rest of the 30% data for testing. Figure 2 shows the outcome of frequency count of correlated or co-occurred words in the review summary. Figure 3 shows the plot of correlated reviews in the training dataset.

The Box plot of the prediction model clearly shows the overall rating of each categorical topics/aspects as shown in Fig 4. If the inverse prediction is greater than zero, it classifies such reviews as positive reviews otherwise the reviews are classified as negative.

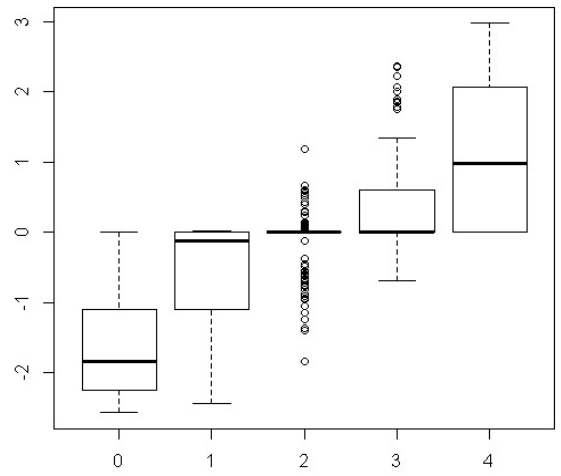

Fig. 4: Box plot of point of scale score of review

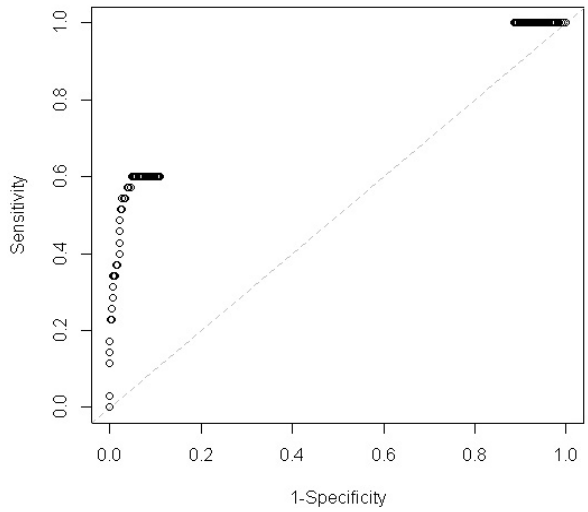

Fig. 5: ROC performance analysis

The ROC curve analysis about a remarkable cut off on inverse prediction that categories the reviews into good or bad are shown in Fig. 5. Eighty six percent of the positive reviews are classified as positive where as eighty one percent of negative reviews are classified as negative. If we increase the unigram into bigram the classification accuracy is also increased significantly.

The final score tells us the classification accuracy of positive score and negative score. The rating of the prediction helps the assessor to assess the overall rating of a review.

CONCLUSION

In this study we propose a probabilistic statistical model of text and aspect mining for extracting rating information for the summarization of e-learning feedback analysis. Our approach takes a collection of review text summary with precision ratings as inputs and discovers each individual learner latent ratings from the summary review. The plots of cowards in the review and co-reviews in the collection helped us to draw and built the sequence flow pattern structure of the individual peer learners reviews. It does not require any metadata or annotated data to discover the latent topics and its corresponding rating. The primary area of future work is to use the semantic correlation of words to the prediction model with the help of ontology system to improve the efficacy of the review analytical system.

REFERENCES

Bespalov, D., B. Bai, Y. Qi and A. Shokoufandeh, 2011. Sentiment classification based on supervised latent n-gram analysis. Proceeding of the 20th ACM Conference on Information and Knowledge Management (CIKM'11). Glasgow, Scotland, UK.

Callan, K.C.T.J., 2004. A language modeling approach to predicting reading difficulty. Proceeding of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL, 2004).

Dankmar, B.H., 1992. Multinomial logistic regression algorithm. Ann. Inst. Statist. Math. , 44(1): 197-200.

Ferrell, G., 2012. A View of the Assessment and Feedback Landscape: Baseline Analysis of Policy and Practice from the JISC Assessment & Feedback programme. Retrieved form: http://www.jisc.ac.uk/media/documents/programmes/elearning/Assessment/JISCAFBaselineReportMay2012.

Granito, A., G.R. Mangione, S. Miranda, F. Orciuoli and P. Ritrovato, 2014. Adaptive feedback improving learningful conversations at workplace. J. e-Learn. Knowl. Soc., 10(1): 63-83.

Hu, M. and B. Liu, 2004. Mining and summarizing customer reviews. Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04). Seattle, Washington, USA.

Jara, M. and H. Mellar, 2010. Quality enhancement for e-learning courses: The role of student feedback. Comput. Educ., 54(3): 709-714.

Lu, B., M. Ott, C. Cardie and B.K. Tsou, 2011. Multi-aspect sentiment analysis with topic models. Proceedings of the IEEE 11th International Conference on Data Mining Workshops (ICDMW '11), pp: 81-88.

McAuley, J., J. Leskovec and D. Jurafsky, 2012. Learning attitudes and attributes from multi-aspect reviews. Proceeding of IEEE 12th International Conference on Data Mining (ICDM, 2012).

Pang, B. and L. Lee, 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceeding of the 43rd Annual Meeting on Association for Computational Linguistics (ACL, 2005), pp: 115-124.

Rabinovich, M. and D.M. Blei, 2014. The inverse regression topic model. Proceeding of the 31st International Conference on Machine Learning. Beijing, China, JMLR: W&CP, Vol. 32.

Roberts, M.E., B.M. Stewart, D. Tingley and E.M. Airoldi, 2013. The structural topic model and applied social science. Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application and Evaluation, 2013.

Skogestad, S. and I. Postlethwaite, 2007. Multivariable feedback control-analysis and design. IEEE Control Systems Magazine, February 2007.

Sun, Y., H. Deng and J. Han, 2012. Probabilistic Models for Text Mining. In: Aggarwal, C.C. and C.X. Zhai (Eds.), Mining Text Data. Springer Science+Business Media, LLC 2012, pp: 259-295.

Taddy, M., 2013. Multinomial inverse regression for text analysis. J. Am. Stat. Assoc., 108: 755-770.

Titov, I. and R. McDonald, 2008. A joint model of text and aspect ratings for sentiment summarization. ACL, 8: 308-316.

Turney, P.D., 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceeding of the 40th annual meeting of the Association for Computational Linguistics (ACL, 2002), pp: 417-424.

Wachsmuth, H., M. Trenkmann, B. Stein and G. Engels, 2014. Modeling review argumentation for robust sentiment analysis. Proceeding of the 25th International Conference on Computational Linguistics: Technical Papers, (COLING, 2014), pp: 553-564.

Wang, H., Y. Lu and C. Zhai, 2010. Latent aspect rating analysis on review text data: A rating regression approach. Proceeding of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'10). Washington, DC, USA.