Asian Journal of Multidisciplinary Research (AJMR)

E-ISSN: 2395-1702 P-ISSN: 2395-0382
Volume 01- Issue 01-, pp-10-17

Research Paper Open Access

AUTOMATED SUTDENT FEEDBACK CONTENT ANALYSIS USING STASTICAL METHODS

B Magesh[1], K Balaji [2].
M.E, Department of Computer Science
Arunai Engineering College,
Tiruvannamalai, India
Magesheera@gmail.com, balajjee.mecse@gmail.com

Abstract— every year massive amount of feedback is gathered from students regarding subjects and its respective faculty. The amount of time to analyze this data manually is a very tedious and time consuming. Pure manual analysis cannot deal with the ever growing scale of data. Automatic summarization is the process of reducing a text document it substantially reduces the costs of analyzing large collections of text using Text mining is concerned with the text analysis of data for finding patterns and regularities in the student feedback data sets. We propose a novel method using multivariate predictive model for conceptual content analytics based on student 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. This will act as a multi-label classifier in order to identify the classification and prediction of student problems and statistically measuring student preferences by analyzing qualitative descriptive gathering reviews.

Keywords—- Student feedback, logistic regression, text mining, review analytics.

1. Introduction

I. The Advanced technology has resulted in large storage of data. These large masses of data consist of some hidden information of strategic importance which can be used for future analysis with effective decision making They are two important types of data analysis methods are classification and predictions Today, the amount of electronically stored documents and, more general, information items increases dramatically. The growth of the web can be seen as an expanding public digital library collection. Online digital information extends far beyond the web and its publicly available information. Vast amounts of information are private and are of interest to local communities, such as the records of customers of a business. This information is mainly text. There are estimates that about 85% of business-relevant information originates in unstructured form. Although its primary purpose is recordkeeping an automated analysis might be desirable to find patterns in the stored records. Therefore, methods for automated text processing can be employed to deal with the information overload of modern communication systems and access the hidden patterns, e.g. by information extraction and document summarization (. Analogous to data mining on structured data, text mining also finds patterns and trends in information, samples that are far less structured but have greater immediate utility for users.

To make a decision on any matter, we always ask the other people “what they think”. The faculty is a building block of any educational institution and students are the ambassador of the teachers. Institutes have different parameter to check the performance of their faculty to enhance the educational quality. One is to take feedback from students about teaching faculty. Universities conduct the online teacher performance survey. Students give their feedback (comments) in textual free format to express their reviews. The data set used in this research has been educational student feedback data to extract the frequently commented features along with their opening words. Three level sentiment classifications (document, sentence and feature level) exist to summarize customer reviews. The objective of analysis and visualization of data is to highlight useful information and support decision making and To finding prediction and forecasting from driven data set by statistical approached The Data analysis and visualization is Automated so easily Understanding data set to take good decision making and Measuring student preferences by analyzing qualitative feedbacks. it reduce manual analysis time and cost paper, 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 student 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 student Reviews 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.

II.

the teachers. Institutes have different parameter to check the performance of their faculty to enhance the educational quality. One is to take feedback from students about teaching faculty. Universities conduct the online teacher performance survey. Students give their feedback (comments) in textual free format to express their reviews. The data set used in this research has been educational student feedback data to extract the frequently commented features along with their opening words. Three level sentiment classifications (document, sentence and feature level) exist to summarize customer reviews. The objective of analysis and visualization of data is to highlight useful information and support decision making and To finding prediction and forecasting from driven data set by statistical approached The Data analysis and visualization is Automated so easily Understanding data set to take good decision making and Measuring student preferences by analyzing qualitative feedbacks. it reduce manual analysis time and cost

2.0 The problem definition and related work

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 (Magdalena Jara, 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 (Dr Gill 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(Bo Pang,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(Kevyn,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(Peter D. 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 (HenningWachsmuth, 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 (Dmitriy Bespalov, 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 (Julian McAuley,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 (Maxim Rabinovich, 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(Matt Taddy,2013).

The multinomial logistic regression model becomes specifically attractive leading to a monotonically converging sequence of iterations (Dankmar bi~hning, 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 A., Mangione G.R,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 (Sigurd, 2007). The summarization technique involves different text inputs and which mostly 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 (Minqing Hu and Bing Liu,2004) . The terms in topics are modeled by multinomial distribution; and the observations for a random field are modeled by Gibbs distribution (Yizhou Sun, 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 (Hongning Wang, 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 (Margaret E. Roberts, 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 (Ivan Titov, 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 (Bin Lu_yz . 2011).

3.0 General architecture

The general architecture of the prediction model consists of fivefold subsystem as shown in the fig 1.

1. Learners Community
2. Instructors Community
3. Moderator (Assessor)
4. Multi nominal Inverse Regression Model
5. 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.

3.1 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 lexicalization 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.

3.2 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.
```

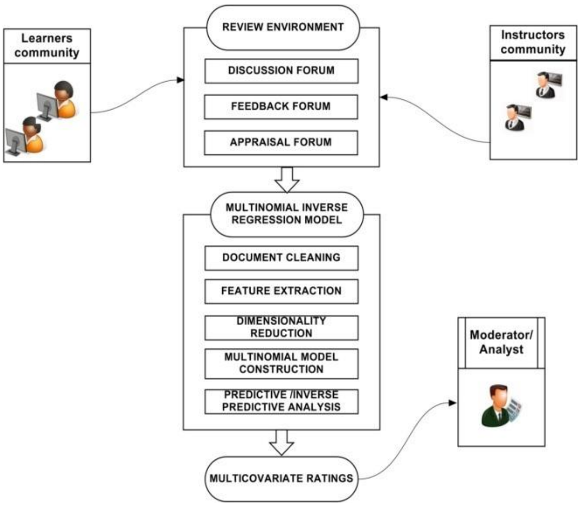


Fig. 1. The general architecture of MPMCA

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

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

(1)

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.

3.2 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.

TOPICAL ASPECTS	**SUMMARY**	**RATINGS(5 POINT SCALE)**
STUDENT 1.Language Problem 2.Availability of contents 3.Illness(self/family) 4.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 1.Syllabus not covered 2.Problems not worked out 3.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**
QUESTION PAPER 1.Tough 2.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 1.Tough 2.Revaluation applied	The valuation seems that it is easy because most of my friends have cleared the semester subjects' have applied photo copy of my answer script.	**2.7**

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 equation (2)

$$P(y\,|\,x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad (2)$$

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

The score is obtained as per the equation (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*.

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

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

$$f(k,i)=\beta_{0,k}+\beta_{1,k}X_{1,i}+\beta_{2,k}X_{2,i}+......\beta_{m,k}X_{m,i} \qquad (5)$$

4.0 experiment and 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. Fig. 2 shows the outcome of frequency count of correlated or co-occurred words in the review summary. Fig. 3 shows the plot of correlated reviews in the training dataset.

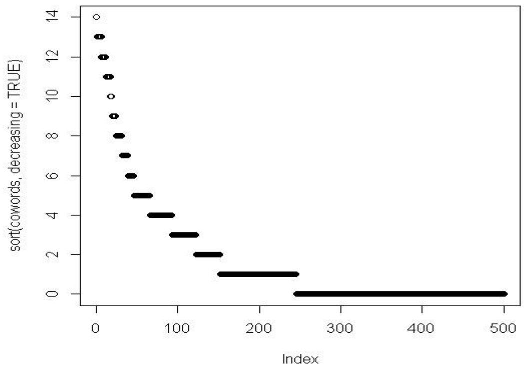
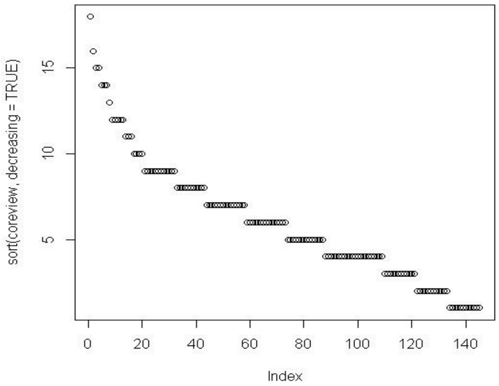


Fig. 2. Plot of words with its co-words



.**Fig. 3.** Plot of words with its Co-words in the overall review

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

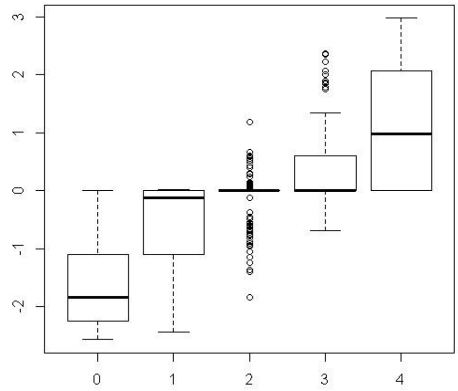


Fig. 4. Box Plot of Point of Scale Score of Review

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

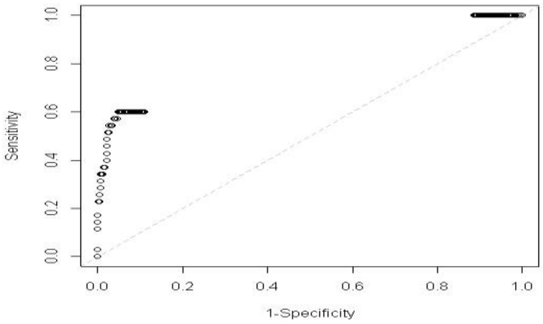


Fig. 5 ROC performance analysis

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.

5. 0 Conclusion

In this paper we propose a probabilistic statistical model of text and aspect mining for extracting rating information for the summarization of student 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 core views 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.

6.0 References

Bin Lu_yz, Myle Ottyx, Claire Cardiey and Benjamin Tsou, "*Multi-aspect Sentiment Analysis with Topic Models"*,ICDMW '11 Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops Pages 81-88,2011.

Bo Pang and Lillian Lee, "*Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales* ",proceedings of the ACL, 2005.

Dankmar bi~hning, "*Multinomial logistic regression algorithm"*, ann. Inst. Statist. Math. Vol. 44, no. 1, 197-200 1992.

Dmitriy Bespalov, Bing Bai, Yanjun Qi, "*Sentiment Classification Based on Supervised Latent n-gram Analysis"*, CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK. ACM 978-1-4503-0717-8/11/10,2011.

Dr Gill Ferrell, "*A view of the Assessment and Feedback Landscape: baseline Analysis of policy and practice From the JISC Assessment & Feedback programme"*, JISC, April 2012.

Granito A., Mangione G.R., Miranda S., Orciuoli F., Ritrovato P. "*Adaptive Feedback Improving Learningful Conversations at Workplace"*, Journal of student Reviews and Knowledge Society, v.10, n.1, 63-83. ISSN: 1826-6223, e-ISSN:1971-8829.2014.

HenningWachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels," *Modeling Review Argumentation for Robust Sentiment Analysis"*, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers,pages 553–564, Dublin, Ireland, August 23-29 ,2014.

Hongning Wang, Yue Lu, Chengxiang Zhai, "*Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach"*, KDD'10, July 25–28, , Washington, DC, USA. 2010.

Ivan Titov, Ryan McDonald, "*A Joint Model of Text and Aspect Ratings for Sentiment Summarization"*, http://zagat.com and http://tripadvisor.com. Association for Computational Linguistics (ACL), 2008.

Julian McAuley, Jure Leskovec, Dan Jurafsky, "*Learning Attitudes and Attributes from Multi-Aspect Reviews"*, IEEE International Conference on Data Mining (ICDM), 2012.

Kevyn Collins-Thompson Jamie Callan, "*A Language Modeling Approach to Predicting Reading Difficulty",*Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004.

Magdalena Jara and Harvey Mellar, "*Quality Enhancement for Student Reviews Courses: The Role of Student Feedback"*, Computers & Education in 2010.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoldi, "*The Structural Topic Model and Applied Social Science"*, Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation. 2013.

Matt Taddy, "*Multinomial Inverse Regression for Text Analysis"*, arXiv:1012.2098v7 [stat.ME] 8 Aug 2013.

Maxim Rabinovich, David M. Blei, "*The Inverse Regression Topic Model"*,Proceedings of the 31 st International Conference on Machine Learning, Beijing, China, JMLR: W&CP volume 32. 2014.

Minqing Hu and Bing Liu, "*Mining and Summarizing Customer Reviews"*, KDD'04, August 22–25, , Seattle, Washington, USA,2004.

Peter D. Turney, "*Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews"*,Proceeding ,Proceedings of the 40th ACM conference, ACL-2002,pages 417-424,2002.

Sigurd skogestad and ian postlethwaite, "*Multivariable feedback control—analysis and design"*, IEEE control systems magazine, february 2007.

Yizhou Sun, "*Probabilistic Models For Text Mining,Mining Text data"*, pp 259-295 ,2012.