CLASSIFICATION OF CUSTOMER COMPLAINTS USING LATENT DIRICHLET ALLOCATION

Leticia H. Anaya, University of North Texas, Denton, TX, USA,
Lanaya@unt.edu, (940) 565-2022
Nicholas Evangelopoulos, University of North Texas, Denton, TX, USA,
Nick.Evangelopoulos@unt.edu, (940) 565-3056

ABSTRACT

In this paper we investigate the potential of employing Latent Dirichlet Allocation (LDA), a relative new text analytics method, as a classifier of unstructured customer comments. In a designed study, LDA classification performance is compared to human classification. The results indicate that humans outperform LDA, but not by far, suggesting LDA as a viable option for automatic processing of large collections of documents.

Keywords: Latent Dirichlet Allocation, Classification, Service Operations.

INTRODUCTION

Customers nowadays have an unprecedented leverage when they engage in a dispute with a product or service provider. Companies are recognizing that this unprecedented ability for negative comments to go viral over the Internet can cause often irreparable business damage in a very short period of time. To counteract the negative impact of these comments, many companies are now using voice of the customer (VOC) technology to integrate the gathering of comments from all possible outlets (e.g., customer relationship management systems, e-mails, customer service records, websites, and social media). Then they are processing these comments through different text analytic techniques to determine underlying insights.

In this paper we investigate the potential of employing a new text analytic method, Latent Dirichlet Allocation (LDA), as a classifier of unstructured customer comments. LDA has been used primarily to extract documents, but its classification performance relative to human has not been adequately researched (Blei 2012) The present paper is a first attempt to classify documents and evaluate LDA classification performance by comparing it to the classification performance of humans. The research question posed in this study is to determine how accurate LDA is in classifying customer comments from a known labeled corpus of textual data and to compare its accuracy performance to the classification done by humans over the same labeled corpus of textual data. This paper is organized as follows. First, an overview of the LDA is presented. This is followed by a research study, designed to address the research questions. Next, the results of the study are presented. After the results are discussed, the paper concludes with a note on contribution and directions for future study.

LATENT DIRICHLET ALLOCATION

Latent Dirichlet Allocation (LDA) was introduced by Blei et al. (2003). In the basic LDA model, a document has a distribution of topics and each topic has a distribution of words and words in one topic can exist in another topic. In LDA, the probability of selecting word w_i from a given topic $z = j$, $P(w_i|z=j)$, follows the Dirichlet multinomial distribution given by:

$$Dir(\alpha_1, \ldots . \alpha_T) = P(P_1, P_2, \ldots P_T / \alpha_1, \ldots . \alpha_T) = \frac{\Gamma \sum_j \alpha_j}{\prod_j \Gamma \alpha_j} \prod_{j=1}^{T} P_j^{\alpha_j - 1} \quad (1)$$

where $0<\alpha_j< \infty$, and T is the number of topics found in a mixture where the proportion of each topic P_j, $j = 1 \ldots T$, ranges from 0 to 1, with the constraint $\sum_{j=1}^{T} P_j = 1$. For simplification purposes, parameter $\alpha_i = \alpha_j = \alpha$, for any i and j. The probability of selecting a topic from a document $P(z=j)$ also follows the Dirichlet multinomial distribution with parameters Dir(β_1, ….., β_T). Again for simplification, parameters $\beta_i = \beta_j = \beta$ are constant for any i and j. Based on the probabilistic topic model, these probabilities are combined to select the probability of finding a word in a given document and this probability is given by:

$$P(w_i) = \sum_{j=1}^{T} P(w_i/z_i = j) P(z_i = j) \quad (2)$$

Obtaining the probability distribution of words in a document is the desired goal for text mining as this allows for the determination of the *best words* (highest probability words) that characterize a document. Blei et al. (2003) expressed the probability of selecting a word from a document, $P(w|\alpha,\beta)$, assuming that the prior $P(z)$ and the posterior probabilities $P(w|z)$ followed multinomial distribution from a mixture of Dirichlet distributions with parameters β and α, respectively, into an intractable expression:

$$p(w|\alpha,\beta) = \frac{\Gamma \sum_i \alpha_i}{\prod_i \Gamma \alpha_i} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{i,j})^{w_n^j} \right) d\theta \quad (3)$$

Because of the intractability properties of these key probabilities presented in expression (3), other approaches in implementing LDA were sought. One of these approaches, Griffiths and Steyvers (2004), utilizes a Markov chain Monte Carlo (MCMC) procedure with Gibbs sampling. The Gibbs sampling procedure begins with a stream of $\boldsymbol{X_o}$ values initialized at time $t = t_o$, which are then incorporated into a function $f_{y|x}(\boldsymbol{y}|\boldsymbol{x})$ to generate another set of random conditioned $\boldsymbol{Y}_i|\boldsymbol{X}_{i-1}$ variables. These $\boldsymbol{Y}_i|\boldsymbol{X}_{i-1}$ variables are then substituted into another $f_{x|y}(x|\boldsymbol{y})$ distribution to generate an updated set of random $\boldsymbol{X}_i$ variables and the process repeats itself iteratively. In this particular MCMC procedure, the state of the system depends only on the previous state and not the past history (Markovian property) and the transition from one state to another occurs randomly (Hogg et al., 2005). In this research, the Griffiths and Steyvers (2004) algorithm based on the Gibbs sampling approach was used with constant α and β parameters of the Dirichlet distribution set to $\alpha = 50/T$, where T is the number of topics, and $\beta = 0.01$.

RESEARCH DESIGN

To answer the proposed research questions, a human/computer classification experiment was conducted for a familiar setting, a pizza restaurant, in two stages. In the first stage, labeled comments were generated by sophomore-level business students via an online survey and in response to specific cues. These specific cues came from an Ishikawa diagram (a cause and effect fishbone diagram) that was developed to determine the causes of customer dissatisfaction with a pizza restaurant and it was based on defining quality as "meeting or exceeding customer's expectations" with a pizza restaurant. The final Ishikawa diagram (Fig. 1) includes 21 specific

causes (subtopics), organized into four main cause categories: Food Condition, Food Processing Methods, Facility, and Customer Service.

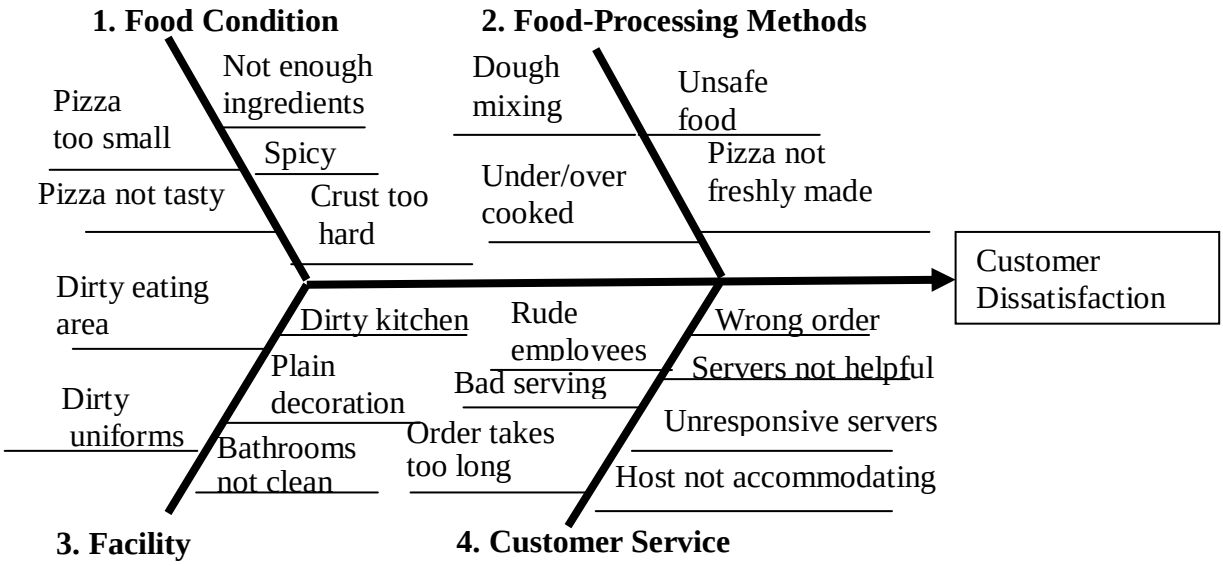

Figure 1. Ishikawa (fishbone) diagram

In the first stage online survey, a complaint was generated from a response to a specific subtopic. This survey consisted of 63 random questions based on 21 specific subtopics to allow for three complaint opportunities per subtopic. The participant was asked to pretend to be a customer and make a complaining statement to include a) an everyday example of the problem cause b) mention of the specific subtopic in the complaint and c) mentioning of the specific major cause category in the complaint. A total of six surveys were administered to over 300 business students yielding more than 6000 complaints. The time spent in completing the survey was recorded and used to determine whether a complaint comment was valid. From this, 1008 complaint comments were deemed valid and actually used in the second stage of the experiment.

In the second stage, thirty six different online survey randomly posted the previous comments and they were presented to human classifiers (MBA students who acted as pizza restaurant managers and who used their own personal skills and knowledge about a pizza restaurant to categorize these comments). Each online survey consisted of 63 complaints, three complaints per subtopic, randomly distributed in the survey. The time to complete the survey was also recorded and used to determine whether the classification was deemed to be valid. Overall, two to four participants evaluated each survey.

For the computer classification stage, the LDA based on Griffiths and Steyvers (2004) topic model with Gibbs Sampling was used. Estimation of the parameters in the word and document topic distributions was performed using Markov chain Monte Carlo (MCMC) simulations. The classification performance measures used to compare both classification methods (human and LDA) were the Macro-F1 and Micro-F1 scores as defined in Tang et al. (2009).

RESULTS

Inter-Rater Reliabilities

The inter-rater reliabilities were computer using Fleiss's Kappa statistic and its modified version, Gwet's AC_1 statistic (Gwett, 2008) for all thirty six surveys. In the psychometrics literature, a Kappa statistic greater than 0.6 is considered to be a "substantial" strength of agreement among the raters. Based on the results, surveys number 2, 10, 19, and 25 were excluded and the remaining surveys were used for the analysis of this research as they had inter-rater reliabilities ranging from 0.644 to .900.

Analysis of the Macro-F1 (MF1) scores for the subtopics resulted in SO1 subtopic (Pizza is not tasty) being an outlier and was deleted from research. The remaining subtopics had Macro-F1 classification scores of 0.463 to 0.953, with an average value of 0.765. The Micro-F1 score for the subtopics was 0.769 and the accuracy rate was 0.764. The three subtopics that humans were able to categorize the most are S11, S13, 14 and the subtopics that humans were able to categorize the least are subtopics S07, S08, S16. These details of these subtopics are listed in Table 1.

Table 1. Unique Human Classification Subtopics Macro-F1 Scores

Easier to Categorize Subtopics
S11=Facility: Employee uniforms are dirty (0.952).
S13=Facility: Environment is just too plain, too undecorated. (0.953).
S14=Facility: Bathrooms are not clean. Score (0.953).
Difficult to Categorize Subtopics
S07=Food Processing Methods: Pizza was not safe to eat(possible contamination) (0.463).
S08=Food Processing Methods: Pizza does not look freshly made (cold pizza) (0.470).
S1=Customer Service: When I need help, it takes too long for the server to notice (0.502).

In the analysis of the main categories, the categorization scores improved. This indicated that humans could categorize better at a higher level of abstraction. The results are shown in Table 2.

Table 2. Main Categories Human Classification Scores

M1 Macro-F1 Score	M2 Macro-F1 Score	M3 Macro-F1 Score	M4 Macro-F1 Score	Average Macro-F1 Score	Micro-F1 Score	Accuracy
0.913	0.881	0.97	0.97	0.934	0.94	0.938

At a higher abstraction level, the results indicate that humans can categorize better. At this level, humans could classify facility and customer service comments better than they could classify food processing comments.

LDA Topic Identification

The first step in using LDA as a method for topic extraction is to identify the extracted topics. For the main categories, LDA extraction of four topics was performed for 50,000 iterations, α =50/T, where T = number of topics and β = 0.01 and these topics were classified into M1, M2, M3, and M4 categories. The identification of these topics was performed through a topic frequency distribution as well by extracting actual documents with each topic from the LDA algorithm. Just as the words were extracted from the LDA algorithm, the actual labeled documents were extracted and these documents were used to identify these topics.

Once the topics are identified, five trials ("chains", in the MCMC terminology) of LDA were run for classification under 4 topics and five more trials were run for classification under 20 topics. All trials were run for 50,000 iterations, with the Dirichlet parameters set to α = 50/T and β = 0.01. The classification results for the main categories and the subtopics are listed in Table 3.

Table 3. LDA Main Categories and Subtopic Categories Classification Scores

Main Categories			Subtopics		
Average Macro-F1 Score	Micro-F1 Score	Accuracy	Average Macro-F1 Score	Micro-F1 Score	Accuracy
0.835	0.842	0.842	0.583	0.64	0.638

At the Main category level, the accuracy rate for LDA was 0.842, which is less than the accuracy rate of 0.938 that was achieved through human classification. Similarly, for the other classification scores, the average macro-F1 Score of 0.835 and average micro-F1 score of 0.842 for the LDA method are less than the respective average macro-F1 score of .934 and average micro-F1 score of 0.940 for the human classification experiment. At the subtopic level, LDA has an average accuracy rate of 0.638 which is less than the human classification rate of 0.764 at the subtopic level. All the other classification scores indicate the average macro-F1 score of 0.583 and average micro-F1 score of 0.640 are less for the LDA method than the respective average macro-F1 score of 0.765 and average micro-F1 score of 0.769 for the human classification experiment. The results indicate that LDA has difficulty categorizing comment S06, S19 and S10, and found it easier to categorize comments S13, S17, S18, and S20. The description of these subtopics is listed in Table 4.

Table 4. Unique LDA Categorization Subtopics

LDA EASIER TO CLASSIFY SUBTOPICS
S13=Facility: Environment is just too plain, too undecorated.
S17=Customer Service: Server is not helpful while I try to make my order decisions.
S18=Customer Service: Host is not willing to accommodate my seating preference.
S20=Customer Service: Server brings the wrong order (wrong crust, wrong toppings,
LDA DIFFICULT TO CLASSIFY
S06=Food Processing Methods: Pizza is under of overcooked.
S19=Customer Service: My order takes too long to be served.
S10=Facility: The restaurant's entire waiting area looks dirty.

DISCUSSION AND CONCLUSION

The results presented in the previous section indicate that humans classify better at higher abstract levels than at lower abstract levels. In this study, the human classification accuracy rate for the main categories was .938 compared to the subtopic classification accuracy rate of 0.764, which represents a 23% improvement over the classification of the subtopics. All the classification metrics used (Macro-F1 score and Micro-F1 score) support this finding of humans classifying better at higher abstract levels. The results also indicate also that LDA classifies better at higher abstraction levels than at lower abstraction levels. At the higher abstraction level, the accuracy rate for LDA was 0.842 and at the lower abstraction level the accuracy rate was 0.638. This represents a 32% improvement over the classification of the subtopics. All the other classification metrics (Macro-F1 score and Micro-F1) support this finding.

The contributions of this study have strong implications for the improvement of information systems that process customer comments. By employing LDA as a document classifier, the present study demonstrates the method's potential to go beyond its extant topic extraction applications. Nowadays, customers are voicing their comments over products and services through many different channels creating a situation where large volumes of comments need to be processed in a prompt manner. Using a classifier such as LDA can help prioritize comments so that companies can respond to urgent comments faster especially when a company faces manpower limitation issues, when it is critical to maintain customer satisfaction and when a company is involved in crisis situations (e.g. airline snowstorms). Thus, this research contributes to the constant quest of maintaining customer satisfaction through the prompt processing of customer comments.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Blei, D. M. (2012). Probabilistic Topic Models. Communications of the ACM, 55(4), 77-84.

Griffiths T. L., and Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Science of the United States of America, 101, 5228-5235.

Gwet, K. L. (2008). "Variance estimation of nominal-scale inter-rater reliability with random selection of raters," Psychometrika, 73(3), pp. 407-430.

Tang, L., Rajan, S., and Narayanan, V. (2009). Large scale multi-label classification via metaLabeler. International World Wide Web Conference Committee (IW3C2), April 20-24, 2009, Madrid, Spain, 211-220.