Midterm Report: Topic Modeling from User Feedback on Facebook Brand Pages

J. G. Snyder, University of Colorado at Colorado Springs

***Abstract*— There exists a wealth of information available in Facebook posts and comments on brand pages, but in order to extract this information, automated processes must be employed. Manual analysis does not scale to the size of the available information. Topic models have recently been shown to be powerful tools to identify latent text patterns. This paper reports on the progress of building a system using Latent Dirichlet Allocation to analyze Facebook posts and comments in an unsupervised manner. This system shows promise for finding reasonable topics that people are posting about.**

***Index Terms*—Topic Modeling, Latent Dirichlet Allocation, Text Mining, Facebook, Social Network**

I. Introduction

ENGAGING with users on Facebook is proving to be an important part of a business' marketing strategy. Facebook allows businesses and organizations to create a page where they can post messages, pictures, videos, links, and applications. A user on Facebook can "like" a page to show that they support the page, but also to subscribe to the page's posts. The default page for a Facebook user lists all the updates of the user's friends, but this page may also list posts from the pages that a user has liked. Facebook employs some proprietary algorithms to order and show the recent posts and updates by how much the system thinks the user will be interested. Although the exact details of this algorithm are not published, it is known that the more a user interacts with the page, the higher probability that the page's posts will be shown to the user. Interactions include using an app, or liking, sharing, or commenting on a post. Consequently, many businesses run contests or promotions to get users to like their page, then post messages geared toward having users like, share, or comment on the post. For example, Disney recently posted on its page this message:

"Finish the lyric: 'They can sing, they can dance. After all, Miss, this is France. And the dinner here is never second _____ !'"

This post is clearly aimed at getting users to comment on the message. Indeed, this post produced over 6500 comments, 1000 shares and 16,000 likes. Another post around the same time is an advertisement for Disney's Theme Parks. This is showing that Disney is making frequent posts to remind users of their brand and then post a few messages meant as advertisements. The following advertisement with a link to a free iPad app was posted on Disney's page and produced more than 15,000 likes, 1000 shares, and 300 comments:

"The magic of Disneyland is closer than you think..."

Although Disney employees certainly review comments for offensive content, I doubt they are manually analyzing the comments with much rigor. Indeed, in addition to Disney's main page, Disney has many pages for each of its major movies, characters, and theme parks each with millions of fans. There is simply too much volume and not enough value in analyzing the content. Additionally, the purpose of eliciting these comments is often times to simply engage with the page's fans and not to elicit meaningful responses.

In addition to allowing a user to comment, like, and share a page's posts, Facebook allows a page to display messages from any user. These messages appear on the brand page for all to see. These kinds of posts will sometimes share thoughts a user has about the brand. Many pages do not allow users to post random messages on their page because moderating the posts can be expensive; however, some brands use Facebook posts as a customer service channel. Employees of the brand will monitor the Facebook page and respond to questions, or issues. A personalized response from a large company can create brand loyalty and make the user feel that the company cares about them.

This project seeks to extract some useful information available in these comments and posts using automated methods. Manual analysis becomes expensive to scale to the size of the available information.

Clustering responses can be a useful way to analyze the kinds of things people are talking about and interested in. Traditional clustering approaches require labor-intensive training, but topic models [1] have recently been shown to be powerful unsupervised tools to identify latent text patterns. Topic models seek to identify the topics contained in textual data. Looking at topics inherent in text over time can show trends in the data. Topics can be used to organize the textual data into classes. In the domain of Facebook brand pages, topic models can be used to understand the concerns and general feelings of the page's fans.

Latent Dirichlet Allocation (LDA) [2] is becoming a standard tool in topic modeling. LDA is a probabilistic generative model, which requires very little human

intervention to create and classify documents. This paper reports on the progress of using LDA to analyze posts and comments from Facebook pages to understand how it can best be used in organizing and analyzing this data.

II. Related Work

Analyzing comments on Facebook is similar to the work of analyzing open-ended survey responses. Jackson et. al. [3] present methods for tagging and clustering user's responses. Their methods involve having around 10 people manually cluster survey responses, then combine their clusters using MSG and k-means clustering. Finally the researchers decide on the appropriate number of clusters and approve cluster labels. This method creates good rigor and eliminates some bias by not creating a tagging scheme before hand, but is quite labor intensive.

Besides Facebook, there are many other sources for user feedback. Product reviews can be an especially good source of information. Lee et al. [4] analyzed reviews on specific products to extract product features and then detect the sentiment of those features. [5] describes a system to extract keywords from product reviews and then detect the sentiment around those keywords. The keywords are then shown in a treemap and color coded according to sentiment. Zeigler et al. [6] built a similar system, but applied it to online survey responses. Although detecting sentiment can provide important information from customer feedback data, a first step is to find good ways of extracting the product features or concerns people are having. The opinion mining methods here focus more on sentiment analysis and suffer from problems related to synonymy and polysemy. These papers used simple tf-idf metrics to extract key words and phrases from the text. I am arguing that simply understanding the topics can be a large part of understanding a user's concerns in a customer feedback situation. Indeed, Rilof et al. [7] find that "topic-filtering and subjectivity filtering are complimentary." In a study of various opinion mining systems, Pang [8] found that Topic extraction is important for documents containing (1) comparative studies or (2) discussions of various features, aspects, or attributes. Certainly Facebook posts talk about many different features of the brand's services falling under category 2.

Lee et al. [9] studied 4 text mining methods including Latent Semantic Analysis (LSA) which analyzes which words occur frequently together to find key phrases and topics; Probabilistic Latent Semantic Analysis (pLSA) which introduces probability theory to determine one topic for every document; Latent Dirichlet Analysis (LDA) which extends the probabilistic model to allow for one document to contain a set of topics with a probability distribution; and Contextual Topic Modeling (CTM) [10] which extends LDA to use a logistic normal distributions instead of dirishlet. Although CTM has shown better results, it is more computationally expensive to use and no available implementation was found, so this project will use the more simple LDA method.

Hong et al. [11] applied LDA to Twitter messages to see if it could predict the topics twitter users would write about in the future. LDA proved to be useful compared to other methods. A large part of the research was finding the best way to model documents. For example, should the document be defined as each Twitter message or an aggregate of all of a user's messages?

III. Problem Description

LDA seeks to solve the following problem: A topic is defined as a probability assigned to each word in the vocabulary. A document is assumed to contain a number of different topics with different probabilities. Given a set of documents, LDA seeks to generate a list of topics, and also find the probabilities of each topic occurring in a particular document.

Applied to the domain under consideration, this would translate to two different problems: (1) Given a set of comments from a particular post, tag each comment with a set of topics. The tagging would be used to count the number of comments about each topic, and allow the end-user to browse the comments by topic. (2) Given a set of posts on a page, assign a set of topics to each post to view how the volume of posts about a particular topic has changed over time.

IV. Proposed Approach

Task 1: Scrape comments and posts from Facebook. First I will need to generate a corpus from the posts and comments on Facebook. I will hand select a set of posts that would be good candidates for analyzing the comments. Additionally, I will hand select some brand pages to scrape for user's posts and associated comments. Facebook provides APIs to scrape publicly available posts and comments from brand pages.

Task 2: Choose a small portion of the collected corpus to tag as a gold standard in order to evaluate LDA's effectiveness. I will ask a small group of friends and family to manually tag comments and posts and cross validate their decisions.

Task 3: Apply the methods of LDA to the corpus. I have found an open source package, MALLET [12], which implements LDA and many other Text Mining algorithms. Some tweaking of parameters will be needed in this stage (e.g the number of topics to find.)

Task 4: Evaluate the results against the gold standard, and also evaluate the results qualitatively.

V. Evaluation

Evaluation of the results will be done in two ways. First the results will be compared against the hand-tagged gold standard. This evaluation will use standard evaluation metrics from document clustering problems as defined in [13]. These metrics include, purity, normalized mutual information, Rand index, and F measures.

One interesting problem with evaluation is that LDA is able

to extract more meaningful topics when it is given more information to analyze; however, having someone tag 10,000 posts with meaningful tags may be too labor for this study. One solution I have come up with is to use the full corpus to run the algorithm against, then compare that against a manual tagged subset.

In addition to using clustering metrics, the results will be analyzed qualitatively to determine if the results provide good insight into the data, and what lessons can be learned from the experiments.

VI. Schedule

Feb 20 – Finish scraper for Facebook data and identify posts and comments that should be scraped. (Complete)

Mar 5 – Finish the user interface required to build the Gold Standard. (Complete)

Mar 12 – Finish the midterm write-up (Complete)

Mar 26 – Integrate MALLET into the system. Run the LDA on the Corpus (Complete)

Apr 9 – Complete gold standard

Apr 16 – Build evaluation metrics to compare LDA results against the Gold Standard

Apr 30 – Tweak parameters of system to achieve better results, and analyze results.

May 7 – Finish Final Paper.

VII. Preliminary Procedures

To enable the creation of a gold standard, a website was created and can be found here [14]. Users login with their Facebook login, and then are presented with a list of datasets to tag. After clicking on a dataset, all the posts from the dataset are displayed with a textbox next to each one. The user can type free-form tags into the textbox. The tags are saved when the cursor leaves the textbox. Using Facebook for login to this system has the added benefit of generating more Facebook access tokens that can be used for crawling.

For the initial dataset, I chose to use Kohl's Facebook page. Kohl's Facebook page provides a good test case because it allows anyone to post on its page and share his/her experience. It is a reasonably popular page with over 6 million fans and deals with many customer service issues. I downloaded the most recent 10,000 posts along with all the associated comments. This represents 111 days worth of posts, and 3MB worth of textual data.

I tokenized the posts using regular expression word boundaries then removing non-letter characters from within the word. I used the large stopword filter that comes standard with MALLET containing 524 English stopwords. Removing many stopwords is important because finding topics that contain high probabilities for stopwords is not useful. Indeed, running LDA without stopwords produced many topics containing the same words. With the Kohl's page, many topics include the word Kohl's with high probability. This does not add much value because of course the posts have to do with Kohl's.

From the Kohl's page, I noticed many posts were about how much was spent, the percent saved, or contained urls. To capture these features, I tokenized dollar amounts into a token labeled "[money]", urls into "[url]", and percent into "[percent]". I also use "?", and "!" as tokens because these characters show that the person posting is asking a question or showing excitement. Indeed these features appear in some of the topics.

VIII. Preliminary Results

Following are the 10 of the 20 topics resulting from a test run against the 10,000 post corpus from Facebook. This was run using 10,000 iterations with the default settings and parameters. Only the top 20 words from each topic are shown. The words are shown in decreasing probability. I have added topic names to each of the topics.

0 – Coupons – [money] [percent] coupon coupons discount pass extra purchase spend spent saved discounts savings weekend total print percent dollar save

1 – Kohl's Credit Card – kohls cash card charge credit pay purchase account late payment expired amount paid bill full money made charged thing

2 – Animal Rights Boycott – [url] page fur click link facebook selling check stop bunnies boycott rock posted friends slaughtered scream hey stops working

3 – Kohl's Giving Contest – give love giving happy contest good makes feel gift enter save vote life entry heart god part make feeling

4 – Happy Customer – ! great kohls shopping deals glad awesome youre wow fun guys helpful wonderful sounds amazing wait friend hear shopper

5 – Time ??? – [number] day days ago hours post month years months tonight minutes office kohl de matter ur weeks search 2nd

6 – Sales and Clearance – sale price kohls [money] prices buy clearance deal bought stores sales items item store check products big paid retail

7 – Favorite Place to Shop / Black Friday Sales – kohls love shopping shop favorite place line friday sales lol black khols nice shopped vera wang night lots shoppers

8 – Bought Items – find clothes size bought pair shoes jeans buy nice stores dress boots clothing found set shirt jennifer big baby

9 – Like Local Store – store local hope home stores live change feedback real open team forward bring coming trip ill idea visit support

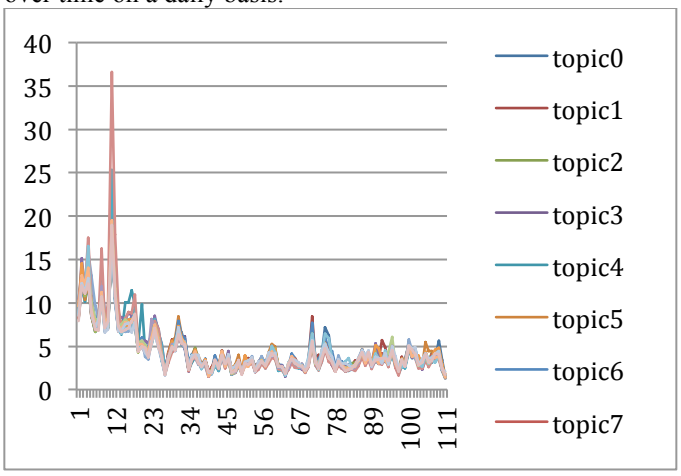
Topic 5 does not appear to be a useful topic. It appears to be a topic of words relating to time. The other topics appear to be useful.

The following table gives a small sampling of the topic / document probabilities for each of the 20 topics. The left-most column is the post id, and the other columns give the percent probability that the post was generated with a given topic. Topic-post combinations with probability greater than or equal to 10 are highlighted in green. This threshold was chosen arbitrarily, but could represent the topics that get "tagged" by the system, just like users would tag posts for the

gold standard. Some documents receive multiple topic assignments with this threshold, and some documents do not receive any topic assignments.

10521	3	2	5	10	7	5	6	4	3	3	4	5	2	7	10	3	6	8	4	5
10520	5	7	4	4	4	7	4	4	4	8	4	10	5	4	5	4	5	8	4	4
10519	3	6	9	6	5	3	13	4	4	6	3	7	4	4	3	7	4	4	4	4
10518	4	4	4	4	4	4	4	4	6	4	4	4	4	19	4	4	4	4	4	4
10517	3	3	6	4	3	8	3	4	8	4	4	9	11	4	3	3	3	8	3	3
10516	8	5	7	3	3	5	5	5	3	6	5	3	3	6	3	5	3	16	3	3
10515	5	3	5	3	3	3	10	3	3	3	5	3	3	6	5	3	9	9	6	5
10514	6	5	3	5	9	3	3	6	3	3	6	3	3	3	3	3	13	3	9	3
10513	5	5	5	5	5	5	5	6	5	5	5	5	6	5	5	8	6	5	5	5
10512	7	3	3	3	2	2	3	4	4	3	3	2	3	3	6	2	24	8	3	4
10511	2	2	7	2	8	11	3	3	8	15	2	5	3	3	3	6	3	5	8	3
10510	3	3	3	3	5	7	5	6	5	3	7	3	3	6	3	8	3	8	8	6
10509	2	17	4	3	3	3	6	7	6	9	2	12	2	3	2	2	3	3	2	2
10508	4	4	4	5	4	4	8	8	4	4	4	4	4	19	4	4	4	4	4	4
10507	4	4	6	4	4	4	4	4	4	6	4	8	6	13	4	4	4	4	4	4

Originally, I thought that looking at topic strengths over time would be useful to see the topics that are important to the people posting on Facebook. My first approach was to simply sum the topic probabilities for all the posts that occurred on a day, then plot the sums of each topic over time. This did not prove useful other than showing the frequency of posting over time. This is confusing because topic 18 is about Christmas. Surely a topic about Christmas would be more popular in December. The following graph shows the topic probabilities over time on a daily basis.



IX. Outstanding Challenges

To summarize the progress thus far, the following challenges still need to be overcome:

1. A method of running the algorithm against a large dataset while only requiring a small dataset for a gold standard
2. How to determine what the threshold should be for saying a post contains a topic.
3. How many topics should be found.
4. Topic-document probabilities seem too uniform.
5. It appears more words should be tagged as stopwords. For example “ur” is used frequently for “you are” and should be considered a stopword.

X. Proposed Solutions

1. I think the solution to this problem is to show performance metrics of the system in two sets. One using the same posts used for the Gold standard, and one using all the posts in the corpus. In either case only tags on the posts used for the gold standard would be compared.
2. Running the LDA calculation and associated metrics against the gold standard can be used to tune the threshold
3. Looking at the Gold Standard would be a good place to start to determine the number of topics.
4. LDA has a parameter, alpha, which can determine the uniformity of the topic document assignments. Basically a lower alpha assigns each document to fewer topics.
5. Perhaps using a tf-idf calculation would be useful here to find more stopwords automatically.

XI. Conclusion

I have proposed a system to automatically group Facebook posts and comments into groups so that patterns and trends can be drawn from the data. Preliminary results were shown describing problems and proposed solutions. This system shows promise for finding reasonable topics that people are posting about.

References

[1] D. Blei and J. Lafferty, “Topic Models,” Text Mining: Theory and Applications, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” The Journal of Machine Learning Research, 2003.

[3] K. M. Jackson, and W. Trochim, “Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses,” Organizational Research Methods, 2002.

[4] D. Lee, O. Jeong, and S Lee. 2008. “Opinion mining of customer feedback data on the web,” In Proceedings of the 2nd international conference on Ubiquitous information management and communication (ICUIMC '08). 2008.

[5] M. Gamon et al., “Pulse: Mining Customer Opinions from Free Text,” Advances in Intelligent Data Analysis VI, Springer Berlin / Heidelberge, 2005, pp. 741.

[6] C. N. Ziegler, M. Skubacz, and M. Viermetz, "Mining and Exploring Unstructured Customer Feedback Data Using Language Models and Treemap Visualizations," Web Intelligence and Intelligent Agent Technology, 2008.

[7] E. Riloff, J. Wiebe, and W. Phillips, “Exploiting subjectivity classification to improve information extraction,” in Proceedings of AAAI, pp. 1106–1111, 2005.

[8] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” Foundations and Trends in Information Retrieval archive. Volume 2 Issue 1-2, 2008

[9] S. Lee, J. Baker, J. Song, and J. C. Wetherbe, "An Empirical Comparison of Four Text Mining Methods," System Sciences (HICSS), 2010 43rd Hawaii International Conference on, 2010.

[10] D. M. Blei, and J. D. Lafferty, “A Correlated Topic Model of Science,” The Annals of Applied Statistics, 1 pp 17-35. (2007).

[11] L. Hong, and B. D. Davison, “Empirical study of topic modeling in Twitter,” In Proceedings of the First Workshop on Social Media Analytics (SOMA '10), 2010.

[12] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu. 2002.
[13] C. D. Manning et al., “Flat Clustering,” Introduction to Information Retrieval, p327. 2008.
[14] http://fb-topic-models.herokuapp.com