Using Latent Dirichlet Allocation for Topic Modelling in Twitter

David Alfred Ostrowski
System Analytics
Research and Innovation Center
Ford Motor Company
dostrows@ford.com

***Abstract*-Due to its predictive nature, Social Media has proved to be an important resource in support of the identification of trends. In Customer Relationship Management there is a need beyond trend identification which includes understanding the topics propagated through Social Networks. In this paper, we explore topic modeling by considering the techniques of Latent Dirichlet Allocation which is a generative probabilistic model for a collection of discrete data. We evaluate this technique from the perspective of classification as well as identification of noteworthy topics as it is applied to a filtered collection of Twitter messages. Experiments show that these methods are effective for the identification of sub-topics as well as to support classification within large-scale corpora.**

1. INTRODUCTION

Social Media has been established as an information source for trend identification [1]. Of equal importance is the determination of the topics behind such trends (i.e. the motivations for each trend) [2]. As a result, a number of areas have grown in interest of summarizing web-based information including topic generation and keyword extraction [3][4].

While Social Media has been leveraged among a number of these areas, it still presents challenges in the interpretation of data. Traditional techniques have not performed well in supporting topic generation and classification, due to non-standard words as well as the overall high amount of noise that is present in such communications. Among the issues are extensive use of symbols, abbreviations, emoticons and non-standard schemas – making this form of communication very difficult to normalize [5].

Latent Dirichlet Allocation (LDA) is a fully generative model for describing the latent topics of documents and is becoming a standard tool in topic modeling [6][7]. LDA models every topic as a distribution over the words of a vocabulary and every document over the sampled topics from a Dirichlet distribution. The words of the documents are drawn from the word distributions of a topic and each topic is drawn for each word from the topic distribution of the document. Inference in an LDA can be accomplished through a number of means including Gibbs Sampling, expectation maximization and expectation propagation [8][9][10]. Popularized in applications for text retrieval, LDA has since then been applied among entity resolution systems, fraud detection in telecommunication systems and image processing [11][12][13].

In this paper we concentrate on the application of topic modeling among specified filtered document collections. Our goal is to provide complementary information among trends for the purpose of Customer Relationship Management (CRM). We consider the application of LDA for the purpose of topic generation by evaluating pre-established topic categories of interest. In the following section we review prior research in this area, Section Three presents our methodology and Section Four our case study. Section Five presents our conclusion.

II. CURRENT RESEARCH

There is currently a substantial amount of work that is leveraging LDA. (Hong et. al.) proposed the use of LDA to derive a topic model within a microblogging (Twitter) environment. By training a topic model on aggregated measures, he demonstrated that topic mixture distributions learned by topic models can support a good set of supplementary features in classification problems, significantly improving overall performance [14]. (Zhao et. al.) applied a Twitter-LDA model in order to discover topics which allowed for a comparison to traditional news media allowing for input into data mining applications [15].

Among classification-based systems, (Biro et.al.) applied LDA to web spam filtering. As applied against a webspam corpus, he was able to demonstrate an 11% improvement in F-measure by a logistic regression-based combination with strong link and content baseline classifiers [16]. (Tian et.al.) was able to compare LDA to Machine Learning algorithms

IEEE ICSC 2015, February 7-9, 2015, Anaheim, California, USA
978-1-4799-7935-6/15/$31.00 ©2015 IEEE

for classification of 41 software collections into domain categories to allow for definitions based on libraries, architectures or programming languages [17]. (Endres et. al.) extended an LDA model to support feature generation, hypothesis generation and statistical modeling of objects in 3D range data, outperforming other unsupervised methods such as hierarchical clustering [18].

LDA has also been applied in online algorithms to support inference techniques over expanding data collections. (Chanini et. al.) utilized inference algorithms for topic models using LDA, incremental Gibbs samplers and particle filters. In their work, they demonstrated how to extend a batch algorithm into a series of online algorithms, maintaining a higher level of flexibility [19]. (Hoffman et. al) developed an online variational Bayes algorithm for LDA. By fitting a 100-topic model to 3.3M articles from Wikipedia, he demonstrated that online LDA techniques could match or surpass those developed with batch techniques [20]. (Wang et. al.) developed an online topic model based on LDA which they applied to 30M microblog postings showing that they were able to reveal interesting topic transitions, such as the work-life rhythm of cities and factors associated with specific problems or complaints.

LDA has been extended in a variety of ways including (Mimno et. al.) who developed a hybrid algorithm for Bayesian topic models that computes the efficiency of sparse Gibbs sampling with the scalability of online stochastic inference [21]. They used their algorithm to analyze a corpus of 1.2M books with thousands of topics. This approach reduced the bias of variational inference and was generalized to many Bayesian hidden-variable models. (McCallum et. al.) proposed a variation of LDA to simultaneously discover groups among the entities and topics within the corresponding text [22]. To support classification efforts, (Maskeri et. al.) applied LDA to extract business topics from program source code for assistance in software maintenance demonstrating that LDA would provide a starting point for documentation efforts [24].

III. METHODOLOGY

LDA is an unsupervised Machine Learning algorithm which identifies latent topic information among large document collections. This technique relies on a "bag of words" approach, which treats each document as a vector of word counts. Each document is represented as a probability distribution over some topics, where each topic is represented as a probability distribution over a number of words. The following generative process is applied for each document in the collection:

1. For each document, select a topic from its distribution over topics.

2. Sample a word from the distribution over the words associated with the chosen topic.

3. The process is repeated for all the words in the document

LDA is based on the hypothesis that a person writing a document has certain topics in mind. The act of topic selection is likened to picking a word with a certain probability from the pool of words of that topic. A whole document can then be represented as a mixture of different topics. Assuming a single author, these topics reflect the persons view of a document and their specific vocabulary. So, each document is represented as a mixture of a fixed number of topics, with topic z receiving weight.

The modeling process of LDA can therefore be described as identifying a mixture of topics for each resource with each topic described by terms following a probability distribution. It can be represented with the following equation (1):

$$P(t_i \mid d) = \sum_{j=1}^{Z} P(t_i \mid z_i = j) P(z_i = j \mid d) \qquad (1)$$

In the above equation, $P(t_i \mid d)$ is the probability of the ith term for a given document d and z_i is the latent topic. Also $P(t_i \mid z_i = j)$ is the probability of t_i within topic j and $P(z_i = j \mid d)$ is the probability of picking a term from topic j in the document. The number of latent topics z has to be defined in advance and allows to adjust the degree of specialization of the latent topics. LDA estimates the topic term distribution P(t|z) and the document topic distribution P(z|d) from an unlabeled corpus of documents using Dirichlet priors for the distributions and a fixed number of topics. Gibbs sampling is one possible approach to this end as it iterates multiple times over each term t_i, in document d_i, and samples a new topic j for the term based on the probability $P(z_i = j \mid t_i, d_i, z_{-i})$ based on equation one until the LDA model parameters converge and is presented as follows (2):

$$P(z_i = j \mid t_i, d_i, z_{-i}) \propto \frac{C^{TZ}_{t,j}}{\sum_t C^{TZ}_{t,j} + T\beta} \frac{C^{TZ}_{t,j}}{\sum_z C^{DZ}_{d_i z} + Z\alpha} \qquad (2)$$

where C^{TZ} maintains a count of all topic-term assignments, C^{DZ} counts the document-topic assignments, z_{-1} represents all topic-term and document-topic assignments except for the current assignment z_i for term t_i, ,and both α and β are the hyperparameters for the Dirichlet priors. Using equation one, the posterior probabilities can be estimated as follows (3)(4):

$$P(t_i \mid z_i = j) = \frac{C_{t,j}^{TZ}}{\sum_t C_{t,j}^{TZ} + T\beta} \quad (3)$$

$$P(z_i = j \mid d_i,) = \frac{C_{t,j}^{TZ}}{\sum_z C_{d,z}^{DZ} + Z\alpha} \quad (4)$$

IV. CASE STUDY

For our experiments, we considered filtered messages from the Twitter firehose that were collected between 1/1/2012 and 12/31/2012. As examined for the purpose of Social CRM, the messages were filtered to include the terms “ford” and “focus”. Upon examination (as matching closely to industry-wide surveys) they were determined to be within four separate categories to the following percentages:

- 45% Blather (jokes, conversational)
- 35% News-driven (events, advertising)
- 15% Interest towards product
- 5% High interest / desire towards purchase

These categories were generalized to characterize the signal as read from the Twitter firehose by means of classification. The three categories of interest were described as:

- Conversational (unrelated content)
- Sales (primarily used car sales)
- Demand (mention of product with significant interest/ desire to purchase)

Starting with a collection of identified sets of 5K messages per category, we considered initial filtering efforts for the removal of stop words and non-alphanumeric characters. This established the starting collection of total words among our three categories (conversation, sales and demand respectively) identified in the following diagram displaying the number of unique words per category (Figure 1).

	category1	category2	category3
total words	41678	37162	39941
unique words	4893	3324	5653

Figure 1. Total and unique words per category.

In our study we applied the LDA algorithm to our established document collection in which we considered the range of three to ten topics. As LDA is unsupervised, there are no assigned document collections, so topics were manually determined among our three initial established categories. Our topics were evaluated by generating clusters determined by a threshold that represents the average of the word probabilities for each topic. In the results below, we considered the precision / recall of our assigned topic categories to our range of generated topics as evaluated from the pre-established threshold.

topics	demand	sales	conv.
	pres/rec	prec/rec	prec./rec
3	.412/.204	.401/.204	.128/.06
4	.417/.112	.477/.296	.328/.184
5	.470/.424	.537/.204	.45/.092
6	.431/.172	.569/.361	.271/.124
7	.520/.412	.662 /.204	.516/.101
8	.470/.424	.489/.188	.45/.092
9	.424/.401	.424/.102	.40/.100
10	.410/.193	.395/.097	.371/.084

Figure 2. Precision / Recall for topic categories

Within our range investigated, the highest topic category for our data sets was seven topics with the precision/ recall performance peaking in all three categories at (.520/.412, .662/.204, .516/.124) for demand, sales and conversational respectively. Overall the precision demonstrated acceptable performance for an unsupervised approach while recall was substantially lower particularly in the conversational category due to a high variability of words. Among the categories, the sales generated the highest levels of precision / recall due to the commonality in the word collection. The conversational was the lowest performing of words due to the highest number of cross-over topics as well as the highest number of unique words.

To evaluate accuracy we applied perplexity, a widely used metric in the topic modeling community. While perplexity is a somewhat indirect measure of predictive performance, it can be shown to be a useful characterization of the predictive qualities of a language model. In our model, we examined the perplexity bounds as a means of evaluating convergence of our topics (figure 3.) . In examination of our results, we were able to see the metric plateau as our tests approached seven topics.

To support keyword/subtopic generation, topics were labelled manually as in the application of clustering.We considered the extraction of only three topics in order to associate to our three starting categories of demand, sales and conversational. Due to our word heterogeneity the most meaningful keywords in each assigned topic categories contained relatively low probabilities than in other published work. A subset of the most relevant keywords per each topic category are presented in figure four. Within the sales category, high probability words also included model years which presented themselves as the most popular cross-cutting topics as they were also identified across the other two assigned categories although at a much lower probability level.

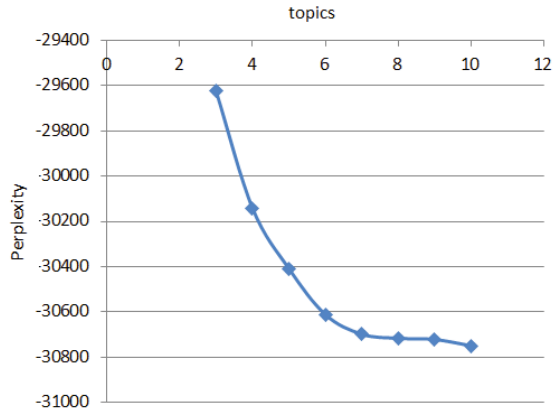

Figure 3. Perplexity measurement as a function of topics.

While some of the most interesting topics upon manual observation were not identified as the highest probability words, some of the general sub-topic themes were captured in our representative collection. Identified topics in the demand category included keywords 'will' and 'going' which supported the immediate intent of buying while 'afford' also provided insight on financial concerns. In the sales category, related terms such as 'used' implied the sales event. Among topics in the conversation category, 'Sirius' represented interest in product features.

demand	
drove	0.01
will	0.02
afford	0.02
milage	0.01
going	0.012

sales	
sales	0.02
used	0.02
auto	0.01
selling	0.03
looking	0.02

conversational	
tonight	0.01
driving	0.01
sirius	0.03
milage	0.03
music	0.02

Figure 4. Sampled Subtopics from LDA model

V. CONCLUSION

In this paper, we have evaluated LDA for the application of topic modeling for the purpose of supporting a greater understanding of trends in Social Media. In our case study, we were able support an unsupervised classification on a set of Twitter information. As applied to our filtered data collection, LDA performed well as an unsupervised model though not surpassing supervised techniques in terms of performance including current application of pure Bayesian-based classifiers. From our results, it has demonstrated potential as a complementary technique that could serve as a means to support the derivation of information about trends as well as assist in identifying new trends.

Future work includes the integration of LDA in conjunction with our supervised techniques. Also we are interested in further experimentation with different thresholds in terms of defining clusters as well as further dividing our initial categories in order to support reduced heterogeneity in our topic models which would allow for further experimentation with the variation of subtopics.

VI. REFERENCES

[1] Gloor, P.A.; Krauss, J.; Nann, S.; Fischbach, K.; Schoder, D.; "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis", Computational Science and Engineering, 2009. CSE '09. International Conference on Volume: 4 pp: 215 – 222

[2]Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?," In Proceedings of the 19th international conference on World wide web, pp. 591-600. ACM, 2010.

[3] Ostrowski, D. A. , "Predictive Semantic Social Media Analysis, IEEE International Conference on Semantic Computing , ICSC , 2011

[4] Baird, Carolyn Heller, and Gautam Parasnis. "From social media to social customer relationship management." Strategy & leadership 39, no. 5 (2011): 30-37.

[5] Kaufmann, Max, and Jugal Kalita. "Syntactic normalization of twitter messages." *International conference on natural language processing, Kharagpur, India*. 2010.

[6] Maskeri, Girish, Santonu Sarkar, and Kenneth Heafield. "Mining business topics in source code using latent dirichlet allocation." *Proceedings of the 1st India software engineering conference*. ACM, 2008.

[7]Hong, Liangjie, and Brian D. Davison. "Empirical study of topic modeling in twitter." *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010.

[8] Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." *Proceedings of the National academy of Sciences of the United States of America* 101.Suppl 1 (2004): 5228-5235.

[9] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[10] Minka, Thomas, and John Lafferty. "Expectation-propagation for the generative aspect model." *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002.

[11]Bhattacharya, Indrajit, and Lise Getoor. "A Latent Dirichlet Model for Unsupervised Entity Resolution." *SDM*. Vol. 5. No. 7. 2006.

[12] Xing, Dongshan, and Mark Girolami. "Employing Latent Dirichlet Allocation for fraud detection in telecommunications." *Pattern Recognition Letters* 28.13 (2007): 1727-1734.

[13] Lienou, Marie, Henri Maitre, and Mihai Datcu. "Semantic annotation of satellite images using latent dirichlet allocation." *Geoscience and Remote Sensing Letters, IEEE* 7.1 (2010): 28-32.

[14] Hong, Liangjie, Ovidiu Dan, and Brian D. Davison. "Predicting popular messages in twitter." *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011.

[15]Zhao, Wayne Xin, et al. "Comparing twitter and traditional media using topic models." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2011. 338-349.

[16]Bíró, István, Jácint Szabó, and András A. Benczúr. "Latent dirichlet allocation in web spam filtering." *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. ACM, 2008.

[17] Tian, Kai, Meghan Revelle, and Denys Poshyvanyk. "Using latent dirichlet allocation for automatic categorization of software." *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*. IEEE, 2009.

[18] Endres, Felix, et al. "Unsupervised discovery of object classes from range data using latent Dirichlet allocation." *Robotics: Science and Systems*. Vol. 2. 2009.

[19]Canini, Kevin R., Lei Shi, and Thomas L. Griffiths. "Online inference of topics with latent Dirichlet allocation." *International conference on artificial intelligence and statistics*. 2009.

[20]] Hoffman, Matthew, Francis R. Bach, and David M. Blei. "Online learning for latent dirichlet allocation." *advances in neural information processing systems*. 2010.

[21] Wang, Yu, Eugene Agichtein, and Michele Benzi. "Tm-lda: efficient online modeling of latent topic transitions in social media." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.
[22] Mimno, David, Matt Hoffman, and David Blei. "Sparse stochastic inference for latent Dirichlet allocation." *arXiv preprint arXiv:1206.6425* (2012).
[23] McCallum, Andrew, Xuerui Wang, and Natasha Mohanty. *Joint group and topic discovery from relations and text*. Springer Berlin Heidelberg, 2007.
[24] Maskeri, Girish, Santonu Sarkar, and Kenneth Heafield. "Mining business topics in source code using latent dirichlet allocation." *Proceedings of the 1st India software engineering conference*. ACM, 2008.]