2015 IEEE 15th International Conference on Data Mining Workshops

OLLDA: A Supervised and Dynamic Topic Mining Framework in Twitter

Shatha Jaradat
Royal Institute of Technology (KTH)
Kista, Sweden
shatha@kth.se

Nima Dokoohaki
Swedish Institute of Computer Science (SICS)
Kista, Sweden
nima@sics.se

Mihhail Matskin
Royal Institute of Technology (KTH)
Kista, Sweden
misha@kth.se

***Abstract*—Analyzing media in real-time is of great importance with social media platforms at the epicenter of crunching, digesting and disseminating content to individuals connected to these platforms. Within this context, topic models, specially LDA, have gained strong momentum due to their scalability, inference power and their compact semantics. Although, state of the art topic models come short in handling streaming large chunks of data arriving dynamically onto the platform, thus hindering their quality of interpretation as well as their adaptability to information overload. As a result, in this manuscript we propose for a labelled and online extension to LDA (OLLDA), which incorporates supervision through external labeling and capability of quickly digesting real-time updates thus making it more adaptive to Twitter and platforms alike. Our proposed extension has capability of handling large quantities of newly arrived documents in a stream, and at the same time, is capable of achieving high topic inference quality given the short and often sloppy text of tweets. Our approach mainly uses an approximate inference technique based on variational inference coupled with a labeled LDA model. We conclude by presenting experiments using a one year crawl of Twitter data that shows significantly improved topical inference as well as temporal user profile classification when compared to state of the art baselines.**

I. Introduction

Topic models have gained strong momentum in recent years for their performance and accuracy in analyzing social media data [1]. This has been most visible in Twitter research with applications on topic classification and user recommendation [2]. Topic modeling is mainly concerned with finding and analyzing the latent relationships between collections of documents in order to perform tasks such as data exploration and prediction. To achieve this, the posterior distribution of the model parameters and the latent variables must be approximated because it is intractable to compute them. Markov Chain Monte-Carlo (MCMC) sampling techniques and variational inference are the two major approaches used for the approximation of posterior inference [3], [4]. Latent Dirichlet Allocation (LDA) [3], is a generative topic model, that has gained attention for its application to mainly supervised topic modeling in social media analytics.

There are a number of shortcomings to LDA, specially when used on Twitter data. LDA is an unsupervised algorithm that models each document as a mixture of topics that can be hard to interpret. This makes it difficult to be directly applied to multi-labeled corpora. A remedy to this problem is applying supervision. In order to achieve this, several modifications have been proposed. One of the most attended approaches is using Labeled LDA (L-LDA) [5]. L-LDA has shown to be a better option than compared to supervised methods such as Supervised LDA [6] and DiscLDA [7], as it doesn't have the constraint of having only a single label associated with each document [5]. However, L-LDA uses Gibbs Sampling as an approximate inference technique, which is one of the MCMC sampling techniques. As such, Gibbs sampling turns out to be less effective than variational inference.

In addition to supervision, another shortcoming of LDA is batch execution, which turns into the streaming nature of content being published on Twitter. To deal with this shortcoming, several online algorithms have been proposed for LDA. One of the most efficient ones is the Online VB for LDA [4]. It can be used to analyze massive collections, including streaming data, with no need for dealing with ephemeral state of documents after they have been processed. However, existing implementation of this approach is also unsupervised, which can have impact on quality of distilled topics.

Variational inference methods can be described as a set of deterministic algorithms that transform the inference problem into an optimization problem. This is done by choosing a simplified distribution that uses a set of free parameters. By solving these parameters and performing optimization steps, the new distribution will be close to the original posterior distribution. It has been shown that this approach is faster than MCMC methods, which calculates the posterior distribution by generating independent samples from the posterior. Existing research have shown that MCMC algorithms can be slow to converge and it is difficult to check their convergence [4], [8], [9], [10]. Within this paper we propose for a scalable and online algorithm that can produce high quality and interpretable topics. Thus, our contributions are: 1) Coupling an online VB for LDA [4] with a Labeled LDA model [5] mainly to improve performance, 2) Using variational inference as an approximate inference technique, which will be enhanced through this work by including supervised learning.

The rest of the paper is organized as follows. Section II provides a background study on the related research. In section III, the proposed algorithm is illustrated in detail. There also, relation to the original online VB for LDA and labeled LDA algorithms has been shown. Section IV presents the results of experimental comparisons, accompanied by an analysis of the algorithms performance and quality. Section V introduces some possible applications of the new algorithm. Finally, the conclusions and the expected future work are presented in section VI.

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.132

II. BACKGROUND

A. Online Learning in Topic Models

Given the importance of dealing with large and streaming data specially in social media analysis, we focus on works that customize topic modeling for streaming scenarios. We hereby focus only on LDA invariants. [11] presents an empirical Bayesian method to incrementally update the model while receiving streaming updates. Suggested by [12], TM-LDA is an LDA invariant that learns the transition parameters among topics, and uses the learnt transition parameters in the topics distribution prediction task. In the same line, OLDA was proposed by [11] , which identifies the emerging topics of text streams and their changes over time. Both [11], [12] are concerned with detecting the patterns in the social behavior. Closely related to our work, there are existing proposals for online versions for LDA, [13], [14]. Both works use Gibbs sampling as the approximate inference distribution method. The work introduced in [13] considered running the sampler on new documents only on the new dataset with each update, whereas [14] proposes incremental Gibbs sampling and particle filters for online inference.

B. Supervised Learning in Topic Models

Existing LDA invariants that incorporate supervision in their process are Labeled LDA [5], Supervised LDA [6], DiscLDA [7]. [5] surveys these algorithms and state that Supervised LDA and DiscLDA both have limitation of associating a document with a single label (topic) only. We argue that this is not suitable for Twitter where each batch of tweets would need multiple labels to be meaningful. In a closely related project, TweetLDA [15], proposed Labeled LDA modeling at the top of Twitter streams. This project highlighted the effectiveness of Labeled-LDA when compared with previous approaches.

C. Coupling Supervision and Online Learning in Topic Models

Given the fact that we propose for an adaptation of supervision to online learning in social media, we focus on two works that influence this work namely Labeled LDA [5] and Online VB for LDA [4]. Labeled LDA [5] incorporates supervision in LDA by constraining the topic model to use topics that correspond to the document observed labels. This has shown to be very effective since instead of assigning the document's words according to their co-occurrence together to latent topics as done by original LDA, it assigns them to labels that have richer meaning, and are based on human classification. The second algorithm [4], is an online Variational Bayes algorithm for LDA, that can handle massive document collections as well as documents arriving in streams. This approach uses variational inference, as an approximate posterior inference algorithm, which is shown to be faster than MCMC [4], [8], [9], [10]. Such consideration makes it an attractive option for applying Bayesian models to large datasets [4]. In our work, we join the positive aspects of two algorithms by modifying the variational distribution used to approximate the posterior distribution in online VB for LDA algorithm, thus constraining the variational parameters to a specific set of meaningful labels. This in turn will influence the assignment of words to topics according to an interpretable classification mechanism.

III. LDA MODELS

As stated previously, within this section we give a brief and technical overview of the original LDA model and some extended LDA models: online VB for LDA, L-LDA and combined online and supervised LDA model, given the context at hand. **Table I** summarizes all the symbols that will be used throughout the text.

TABLE I. LIST OF SYMBOLS

Symbol	Meaning
d	document - sequence of N words
θ	topics' distribution over documents
β	words' distribution over topics
V	Vocabulary used in the model
z_{dn}	topics assignments for the word w with index n in d
w_{dn}	observed word from document d with index n
D	corpus - a collection of M documents
N	number of words in each document
M	number of documents per corpus
K	list of all topics in the corpus
γ	variational parameter mapped to θ
ϕ	variational parameter mapped to z
λ	variational parameter mapped to β
q	approximated variational distribution
p	posterior distribution of LDA
Λ	list of labels assigned to d
T	number of topics assigned to each document

The smoothed Latent Dirichlet Allocation (LDA) model assumes availability of K topics, where each topic is defined as a multinomial distribution over the vocabulary, drawn from a Dirichlet, $\beta_k \sim Dirichlet(\eta)$ distribution. For each document, a distribution over topics θ_d is drawn from $Dirichlet(\alpha)$ where α is a hyperparameter. Then, for each word i in the document, a topic index $z_{di} \in \{1, ..., K\}$ is drawn from the topic weights $z_{di} \sim \theta$, and the observed word w_{di} is drawn from the selected topic $w_{di} \sim \beta\ z_{di}$ [3].

Figure 1 (a) shows the graphical representation of the original model of LDA [3]. The posterior distribution of the hidden variables given the documents is as follows: $p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}$. This distribution can be approximated using different techniques including variational inference.

When variational inference is applied to the approximation of posterior distribution, the LDA model can be transformed into a simpler model as shown in Figure 1 (b). Let us assume this new variational model is named q. In such simplified model, the latent variables θ and β are decoupled by dropping the node w, and removing the edges between z, w and θ. Then the free variational parameters γ, ϕ, and λ are added to the new model to approximate θ, z, and β respectively.

With the new simplified distribution q(z, θ, β), the goal eventually becomes finding the settings of the variational parameters λ, γ, and ϕ that make q close to true posterior distribution p. Thus optimization problem can be solved by minimizing the Kullback-Leibler (KL) divergence between the true posterior p and the variational posterior q. The KL divergence cannot be exactly minimized. However, it is possible to maximize an objective function that is equal up to a constant; the ELBO (evidence lower bound) on the log-likelihood, which shall pertain the same effect as minimizing the KL divergence. To achieve this, Jensen inequality is often used to obtain an

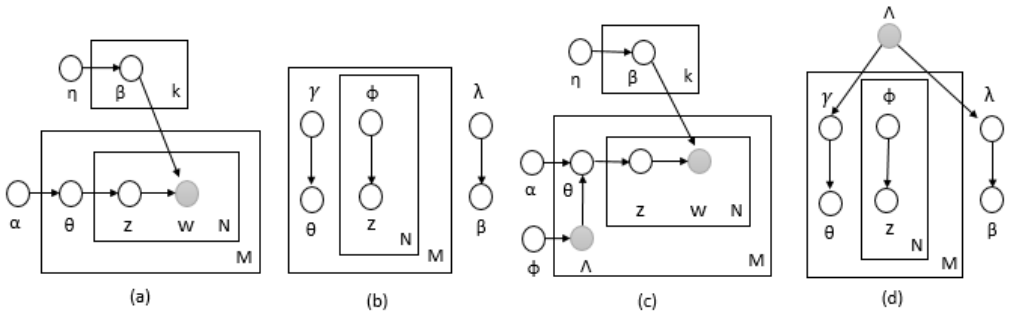

Figure 1: Graphical representations for (a) LDA model - Source [3] (b) variational distribution used to approximate the posterior distribution in LDA - Source [3] (c) Labeled LDA model - Source [5] (d) modified variatoinal distribution used to approximate the posterior distribution in OLLDA

adjustable lower bound on the log likelihood [3], [4]. This is followed by an iterative method, such as variational EM algorithm to iteratively optimize each variational distribution holding the other parameters fixed. In the expectation step, γ and ϕ will be iteratively updated while λ fixed. In the maximization step, λ will be updated given ϕ. Consequently, those updates are guaranteed to converge to a stationary point of the ELBO [3], [4].

A. Online VB for LDA

In online VB for LDA [4], the update step is the one in which λ is updated. The expectation step is executed for the newly observed *t*th vector of words (n_t). In this step, the values of γ_t and ϕ_t are calculated, while holding λ fixed. In this case, the entire corpus is assumed to be consisting of the newly observed document, repeated D times, where D is the number of documents in the batch. The value of $\tilde{\lambda}$ is calculated under such setting. $\tilde{\lambda}$ can be used to update λ in the maximization step, which is computed from the weighted average of its previous value, and the newly calculated value $\tilde{\lambda}$.

Algorithm 1 Online variational Bayes for LDA

Define ρ = $(\tau$ + t$)^k$
Initialize λ randomly
for $t = 0 \infty$ **do**
    E step:
    Initialize γ_{tk} = 1
    repeat:
        Set $\phi_{twk} \propto$ exp $\mathrm{E}_q[log\theta_{tk}]$ + $\mathrm{E}_q[log\beta_{kw}]$
        Set γ_{tk} = α + Σ_w ϕ_{twk} η_{tw}
    until $\frac{1}{K}$ Σ_k |change in $\gamma_{tk}| < 0.00001$
    M step:
    Compute $\tilde{\lambda}_{kw}$ = η + D η_{tw} ϕ_{twk}
    Set λ = (1 - ρ_t) λ + ρ_t $\tilde{\lambda}$
end for

Step one in **Algorithm 1** is used to give weight to $\tilde{\lambda}$, where k $\in (0.5, 1]$ controls the rate at which old values of $\tilde{\lambda}$ are forgotten. The formulas used to calculate $\mathrm{E}_q[log\theta_{tk}]$ and $\mathrm{E}_q[log\beta_{kw}]$ are given as follows:

$\mathrm{E}_q[log\theta_{tk}]$ = Ψ (γ_{dk}) - $\Psi(\Sigma_{i=1}^{K}$ $\gamma_{di})$; $\mathrm{E}_q[log\beta_{kw}]$ = Ψ $(\lambda_{kw})-\Psi(\Sigma_{i=1}^{W}$ $\lambda_{ki})$

Where Ψ is the digamma function.

B. L-LDA

L-LDA, as visualized in **Figure 1 (c)** incorporates supervision in LDA by constraining the topic model to use only the topics that correspond to a document's observed label set. As it is seen from the figure, the topic distributions over documents θ don't depend only on the hyper-parameter α, but also on Λ which is the set of observed labels. This affects the initial assignment of words to topics and topics to documents [5]. In this model, the number of topics can be dynamically assigned according to the number of labels in the corpus. The vocabulary can be composed from the words of the documents. Whereas in online VB for LDA [4], a fixed vocabulary from English dictionary is often used.

C. Online Labeled LDA (OLLDA)

Using concept of constraining, topic distributions per each document are curbed down to a set of observed labels. To do so, we have modified the variational distribution that is used to approximate the posterior distribution in the online labeled LDA model as follows. We have the initial values of the variational parameters γ and λ affected by the set of topics observed for the document Λ , as shown in **Figure 1 (d)**. This in turn affects the probability of both assigning topics to documents as well as words to topics. Certain topics will receive higher probability of being assigned to documents than others during the learning process. This is different than the original LDA model that assigns words to topics only based on their ensemble co-occurrence. Also original LDA model uses a fixed English vocabulary during the assignment process. However, on Twitter most tweets contain hashtags, abbreviations, and names of people which are not part of English vocabulary. This problem is addressed in our project by extracted words from tweets, and assigning them to topics, taking into consideration the prior topics' distribution ,score which we have estimated through the classification service. We believe that this step helps us to achieve more accurate results. As shown later on, when we ran our algorithm with a fixed vocabulary from an English dictionary, the results were not as accurate as the dynamic vocabulary.

Algorithm 2 presents the proposed algorithm. Assume *D* is the number of documents in the corpus. *T* is the number of topics considered from the ground truth for each document, which is a subset from Λ. In our research, we considered the top five topics. The algorithm has been experimented to adopt

Algorithm 2 Online Labeled LDA (OLLDA)

1: $Define\rho$ = $(\tau$ + t$)^k$
2: Initialize λ
3: Initialize γ
4: **for** $d = 0$ to D **do**
5: **for** $t = 0$ to T **do**
6: **if** $SortedTopT(t) \notin K$ **then**
7: Add $SortedTopT(t)$ to K
8: **end if**
9: **end for**
10: **for** $l = 0$ to $WordsRatio$ **do**
11: Add d(l) to V
12: **end for**
13: **end for**
14: **for** $d = 0$ to D **do**
15: **for** $t = 0$ to T **do**
16: $P_{dt} = \left\lceil \frac{|d| * SortedTopT_w}{W_d} \right\rceil$
17: **for** $i = P_{dts}$ to P_{dte} **do**
18: λ_{dk} ++
19: γ_{dkwi} ++
20: **end for**
21: **end for**
22: **end for**
23: Normalize data
24: **for** $t = 0$ to ∞ **do**
25: E step:
26: Initialize γ_{tk} = 1
27: **repeat**:
28: Set $\phi_{twk} \propto$ exp E$_q[log\theta_{tk}]$ + E$_q[log\beta_{kw}]$
29: Set γ_{tk} = α + Σ_w ϕ_{twk} η_{tw}
30: **until** $\frac{1}{K}$ Σ_k |change in $\gamma_{tk}| < 0.00001$
31: M step:
32: Compute $\tilde{\lambda}_{kw}$ = η + D η_{tw} ϕ_{twk}
33: Set λ = (1 - ρ_t) λ + ρ_t $\tilde{\lambda}$
34: **end for**

top three and top five topics, and it has been observed that top five topics result in better topical inference output. In our proposed solution, ***SortedTopT*** is the set of top T sorted labels associated with each document and provided by the ground truth. At the same time, ***WordsRatio*** is the percentage of words allowed to be taken from each document. Each document is partitioned according to the associated topics weights, where P_{dt} represents the size of the partition in document d that will be assigned to topic t, and P_{dts} indicates the starting index of the partition, while P_{dte} is an indication of the ending index. To finalize, main steps of the algorithm can be summarized as follows:

1) Initialization of λ and γ with random values.
2) Topics in corpus are decided from labels associated with documents.
3) Vocabulary is built from the words of documents.
4) For each document, a number of iterations are executed to add value for each top topic in the document, and to assign words to that topic, according to that topic's weight.
5) Rest of the steps match the original algorithm.

In steps 14 to 22, we use the topic's weight that is provided by the classifier to decide the percentage of the document's length that will be modified. We calculate the size of each partition, which is used to increase the importance of a certain topic in a document and certain words belonging to a topic according to the following rule: $P_{dt} = \left\lceil \frac{|d| * t_w}{W_d} \right\rceil$,where P_{dt} is the size of partition per each topic t, d is the document under study, and t_w is the topic's weight in the document. W_d is the total of all topics weights associated to document d. The value of a certain topic in a document, and certain words related to that topic will be incrementally updated during the learning process. This affects their final classification in the context of the whole corpus at the end of the process. **Table II** illustrates the sub-datasets with the number of tweets per each month.

TABLE II. NUMBER OF TWEETS PER EACH MONTH FOR THE LIST OF USERS IN THE SAMPLE

Mar	April	May	June	July	Aug	Sept
44625	83875	159971	109373	94497	52479	186856

IV. EXPERIMENTS

A. Dataset

Dataset used in this research was gathered by [16] using Twitter Streaming API[1] from February 2014 until mid of October 2014. Some filters were applied while gathering the data such as location which was set to Sweden. The total number of tweets is around 7 millions, which are mapped to 471,086 users. The chosen sample is 731,676 tweets corresponding to 3061 users. The sample size was reduced due to limitations in the number of allowed requests to consume the classification service, which will be described later. Due to focus on temporal variations of topics, dataset was partitioned. Multiple partitions were constructed for the users, including their tweets from March 2014 until end of September 2014. Each partition represents tweets of all those users during a specific month. The reason for this partitioning is the interest of this work in training the algorithms on the datasets to notice the transitions of topics from month to month. We were also interested in checking the ability of the new algorithm to detect topics in an accurate way, when compared to the ground truth. Within the course of the experiments the proposed algorithm is compared with the online VB for LDA, as well as L-LDA. For the purpose of comparison, it was required to apply an update mechanism for L-LDA. According to [17], adding a certain percentage of documents that were used in a previous iteration to the new iteration, satisfies the update step in L-LDA. Multiple experiments were conducted, as will be illustrated in the following sections. Experiments have been repeated for different values of the hyper-parameter alpha : 0.1, 0.25, 0.5, 1.0, 1.25, 1.5, 1.75 and 2.0. We observed that in general OLLDA achieves better results for different settings of alpha.

B. Experimental Pipeline

To adapt the algorithm to the Twitter data at hand, we have proposed for the pipeline visualized in **Figure 2**. The goal of this procedure is to annotate tweets of each user with the appropriate labels produced by the classifier. To achieve this, tweets are pooled according to the author, following

[1]https://dev.twitter.com/streaming/overview

the "Author-wise pooling" concept described in [18]. Then stop words are removed from those documents. For our experiment, a translator API was used to translate all the tweets from Swedish to English. For this purpose, **Microsoft translator API**[2] was used, due to its competitive quality of translation. For labels' assignment to user profiles, **Textwise Classifier**[3] was used to provide the ground truth for the sample dataset. This classifier was chosen due to its accuracy in text classification, when compared to other classifiers, such as Alchemy API[4]. For the purpose of this research, we have used an advanced classification service. We could have used some natural language processing tools such as TweetNLP [5] for preprocessing steps and enhancing the analysis results.

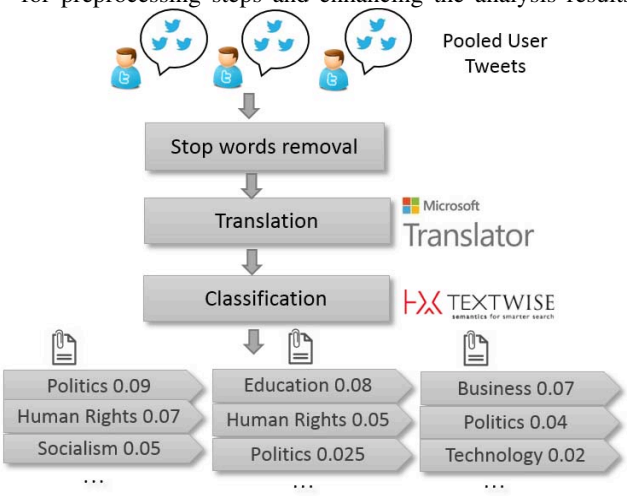

Figure 2: Steps performed before running OLLDA

C. Topical Inference Accuracy

To compare the quality of topical inference, a number of experiments were customized in several settings for each month. To do so, we divided the data in each execution to a 20% for testing and 80% for validation. This setup was repeated for each consecutive month incrementally. As example, in the first iteration, each algorithm was trained on March partition, and subsequently April partition was used for testing and validation. During the second iteration, March and April were used for training, and the testing was done on May partition, etc. This setup was intentionally done to compare the accuracy of detecting topics of each month with previous months. Another reason was to figure the transitions in the inferred topics from one month to another. After running OLLDA and L-LDA, the non-classified profiles were compared with the ground truth (which is the data provided by the classifier), and the cosine similarity was estimated for each month. **Figure 3** compares the percentage of similarity between the topics provided by the classifier, which reflects the user's preferences and the detected topics by the algorithms. The results show that OLLDA has higher cosine similarity values when compared to L-LDA. Experiments were executed for different alphas, and in all results OLLDA was better than

[2]https://www.microsoft.com/translator/api.aspx
[3]http://textwise.com/
[4]http://www.alchemyapi.com/
[5]http://www.ark.cs.cmu.edu/TweetNLP/

TABLE III. PRECISION VALUES FOR DIFFERENT MONTHS

	April	May	June	July	Aug	Sept
OLLDA	0.995	0.992	**1.000**	0.992	0.992	0.997
LLDA	0.986	0.977	**0.982**	0.987	0.992	0.923

TABLE IV. RECALL VALUES FOR DIFFERENT MONTHS

	April	May	June	July	Aug	Sept
OLLDA	0.202	0.201	**0.203**	0.201	0.201	0.202
LLDA	0.199	0.198	**0.199**	0.200	0.201	0.187

counterparts. Although L-LDA considers top topics for each profile to augment their chance of representing the document, it doesn't consider the order and the proportion of each topic for respective profile. This is while, in our proposed approach, we consider those factors while incrementing the value of top topics in each profile during the learning process.

To measure accuracy of retrieval, Precision, Recall scores were estimated for the months of studies. Respective results are illustrated in tables III and IV rounded to three digits. Within our work we defined the precision as the proportion of profiles having correctly inferred topics, over the total count of profiles (test). Using the same reasoning, recall was estimated as the proportion of profiles having correctly inferred topics over the total count of all profiles (test and validation). Results show very strong precision values for assigning correct topics to each profile under study.

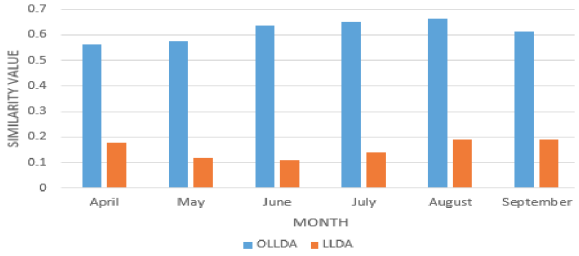

Figure 3: Comparison of cosine similarity estimates between OLLDA and LLDA, with the ground truth for all months.

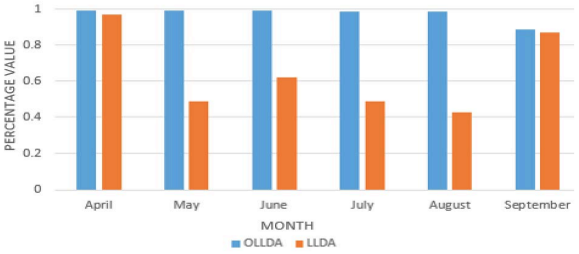

Figure 4: Comparing correctly classified profiles for top (one) topic between OLLDA and LLDA with the ground truth for all months.

Figure 4 plots a comparison of the percentage of correctly classified profiles after running both algorithms for the same testing scenarios and for the top inferred topic. The percentage was calculated based on the comparison between the inferred top topic and the top topic provided by the classifier, assuming that the classifier reflects the user's top favorite topics. The results show better performance for OLLDA when compared to L-LDA.Table V presents the top words that are associated with some of the detected topics in our sample. The majority of users, who are included in the sample, have political and business interests. Figure 5 visualizes a word cloud for a subset

TABLE V. TOP WORDS PER TOPICS

Politics	Economic	Territorial Disputes	Political Science
migpol	heritage	democrat	immigrant
svpol	budget	debate	sd
debate	revenue	party	eupol
Socialism	Human Rights	Education	Business
mingle	financial	belnar	attefall
talent	member	talk	budget
budget	work	val	european

Figure 5: Word cloud generated for a subset of the sample for users with political interests

of users with political interests.

V. OLLDA APPLICATIONS

Twitter data has been used extensively in predicting and explaining a variety of real life events. One of the interesting applications in this area is analyzing political tweets. Dokoohaki et al. [16], applied a link prediction approach on Twitter data that was gathered along the time-line of European and general Swedish elections during 2014, in which they highlighted the correlation between the density of politicians' conversations and their popularity, that is directly reflected in the estimated vote outcomes. A possible direction of OLLDA could be augmenting such analysis by topic modelling to extract and generate debate topics dynamically thus understanding the major themes in politicians' conversations.

OLLDA can be used to calculate the similarity metrics between users in social networks, which has a direct application for this algorithm in recommendation engines. Similarity in topics and interests can be considered as one of the factors that decide the trust level between users. Therefore, OLLDA can be applied in a framework of tweets and friends recommendation as an example. Another possible application is trend detection which might require enhancements of the algorithm, to distinguish the trending topics. Currently OLLDA is tested in Twitter context, it will be interesting to test it in another social network or system to verify its ability in producing better detection and inference of topics. Due to time limitations, OLLDA is still not applied in an extended application, which will imply its testing in a runtime scenario as well.

VI. CONCLUSIONS

This manuscript set forth the Online Labeled LDA (OLLDA), an enhanced online algorithm for topic modeling on Twitter. In our approach, we modified the approximate inference technique used in the selected online version of LDA, by adding a constraint on the variational parameters used in the approximation. This constraint changed the approach to be supervised, making it useful for streaming scenarios while maintaining accuracy. We conducted experiments on a one year Twitter data, with results demonstrating enhanced performance and topical inference quality of the new algorithm, when compared to the state of the art baselines. For future work, we plan to explore applications of framework in recommendation systems.

REFERENCES

[1] M. Pennacchiotti and S. Gurumurthy, "Investigating topic models for social media user recommendation," in *Proceedings of the 20th International Conference Companion on World Wide Web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 101–102.

[2] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi, "Inferring user interests in the twitter social network," in *Proceedings of the 8th ACM Conference on Recommender Systems*, ser. RecSys '14. New York, NY, USA: ACM, 2014, pp. 357–360.

[3] M. Jordan, D. Blei, and A. Y. Ng., "Latent dirichlet allocation," *Machine Learning Research*, vol. 3, pp. 993–1022, Jan. 2003.

[4] M. Homan, D. Blei, and F. Bach, "Online learning for latent dirichlet allocation," *Neural Information Processing Systems*, vol. 23, pp. 856–864, 2010.

[5] R. Nallapati, C. Manning, D. Ramage, and D. Hall., "Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, Aug. 2009.

[6] D. Blei and J. McAuliffe, "Supervised topic models," *NIPS*, 2007.

[7] F. S. Lacoste-Julien and M. Jordan, "Disc lda: Discriminative learning for dimensionality reduction and classification," *NIPS*, Dec. 2008.

[8] M. Jordan and D. Blei, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121– 144, 2006.

[9] H. Attias, *A variational Bayesian framework for graphical models*. Cambridge, MA: Advances in Neural Information Processing Systems 12, MIT Press, 2000.

[10] T. Jakkkola, L. Saul, M. Jordan, and Z. Ghahramani, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999.

[11] C. Domeniconi, L. AlSumait, and D. Barbara, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, pp. 3–12.

[12] M. Benzi, Y. Wang, and E. Agichtein, "Tm-lda: efficient online modeling of latent topic transitions in social media," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and data mining*, Beijing, China, Aug. 2012, pp. 123–131.

[13] C. Teo, J. Eisenstein, A. Smola, E. Xing, A. Ahmed, and Q. Ho, "Online inference for the infinite topic-cluster model: Storylines from streaming text," *Artificial Intelligence and Statistics AISTATS 15*, pp. 101–109, 2011.

[14] L. Shi, K. Canini, and T. Griths, "Online inference of topics with latent dirichlet allocation," in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009.

[15] D. Quercia, H. Askham, and J. Crowcroft, "Tweetlda: supervised topic classification and link prediction in twitter," in *Proceedings of the ACM Web Science*, NY, 2012, pp. 373–376.

[16] N. Dokoohaki, F. Zikou, D. Gillblad, and M. Matskin, "Predicting swedish elections using twitter: A case for stochastic link structure analysis (new)," in *the 5th workshop on Social Network Analysis in Applications (SNAA2015), collocated with IEEE/ACM ASONAM 2015*, Aug. 2015.

[17] A. McCallum, L. Yao, and D. Mimno, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, Paris, France, 2009.

[18] W. Buntine, L. Xie, R. Mehrotra, and S. Sanner, "Improving lda topic models for microblogs via tweet pooling and automatic labelling," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, Jul. 2013.