COMS 4772 Advanced Machine Learning Final Project: MedLDA on Twitter User Classification

Xuejun Wang
UNI: xw2355

ABSTRACT

This project focuses on exploring the possible applications of a recent topic model on user classification on social media, specifically Twitter. The topic model, maximum entropy discrimination latent Dirichlet allocation (MedLDA), integrates the mechanism of generative approach (Latent Dirichlet Allocation), and discriminative approach (Maximum Entropy Discrimination). As was proposed and proved in [1], this model produce more discriminative latent topical representation compared to just using Bayesian topic model, and therefore are more suitable for classification tasks.

1. INTRODUCTION AND RELATED WORK

Applying machine learning approaches in analyzing attributes of Twitter users has been explored in [5] and [6]. And attributes like user profile, tweeting behavior (types of information like hashtags or URLs that are included in tweets), linguistic contents, and user's social network (followers and following attributes). However, compared to analyzing tweets' linguistic contents, analyzing URLs and multimedia contents or social connections that are associated with users requires more complex web crawling, text and graph analysis, and perhaps a comprehensive model to represent these attributes for prediction.

The machine learning model adopted in [5] originates from the Latent Dirichlet Allocation model proposed in [2]. Each user is represented by a multinomial distribution over topics drawn from a Dirichlet prior. Each topic is represented by a multinomial distribution drawn from another Dirichlet prior. Both prior distributions along with topic assignments according to topic per user prior then form a joint distribution for the corpus. Then approximate inference is to be performed to maximize the marginal likelyhood for the joint distribution. Classic Support Vector Machine is used in [6] to classify and predict user attributes, exploiting its maximum-margin property. However both models are only applied on binary classification tasks, where user attributes are studied only on "black or white" basis, which is far from being applicable for real life scenarios.

Two sets of user attributes are studied in this projects. First set is user's political orientation that falls in the following 8 types: Libertarian, Liberal, Progressive, Moderate, Tea Party, Conservative, Anarchist and Revolutionary. They share similarities in hashtags and words as is shown in experiments, but also differentiate from each other in political interests, which is expected to result in different corpus distribution. Another set is 4 classes of rather heterogeneous personal interests: Technology, Music, Sports and Gaming. They are expected to have low similarity in corpus distribution. User data are extracted using Twitter API and constructed in the required format topic model package provided by petuum[9].

2. MEDLDA: MAXIMUM MARGIN SUPERVISED TOPIC MODELS FOR CLASSIFICATION

The purpose of MedLDA classification model is to learn a supervised topic model provided documents/users labels are available in training stage, such that it can be used for predicting labels for new documents/users. The discrete response variable y for the supervised model can take values from a finite set. The latent topic assignments of all words in a document are given for classification: $\mathbf{z} \triangleq \{z_1 \dots z_N\}$. The effective latent linear discriminant function can be defined as:

$$F(y; \mathbf{w}) = \mathbb{E}[F(y, \mathbf{z}, \eta; \mathbf{w}] = \mathbb{E}[\eta_y^T f(y, \bar{\mathbf{z}}) \mid \alpha, \beta, \mathbf{w}]$$

where $\bar{\mathbf{z}} \triangleq 1/N \sum_n z_n$, w denotes all words in corpus, α being the topic mixing prior, β being the $M \times K$ topic distribution matrix(M being the total number of words), η_y is a class-specific K-dimensional vector associated with class y(K being the number of topics); η obtained by stacking the elements of

η_y. Then the prediction rule for such model is:

$\hat{y} = argmax_y F(y; \mathbf{w}) = argmax_y \mathbb{E}[\eta_y^T f(y, \bar{\mathbf{z}}) \mid \alpha, \beta, \mathbf{w}]$

The goal then becomes to obtain optimal set of parameters (α, β) and distribution $q(\eta)$ such that the above expectation is maximized. Latent Dirichlet Allocation model is therefore adopted to model the likelihood of documents. Document labels Y are used in max-margin learning instead of providing feedback as opposed to the purpose in supervised LDA, and the integrated problem of discovering latent topical representations and learning a distribution of classifiers is:

$$min\mathcal{L}^u(q; \alpha, \beta + KL(q(\eta) \parallel p_0(\eta) + \frac{C}{C}\sum_{d=1}^{D}\xi_d$$

$$\forall d, y, s.t. : \begin{cases} \mathbb{E}[\eta_y^T \Delta f_d(y, \bar{\mathbf{z}})] \geq \Delta l_d(y) - \xi_d \\ \xi_d \geq 0 \end{cases}$$

where the variational bound $\mathcal{L}^u$ is defined and deduced to have the below lower bound in LDA:

$$\mathcal{L}^u \triangleq -\mathbb{E}_q[\log p(\{\theta_d, \mathbf{z}_d\}, \mathbf{W} \mid \alpha, \beta)] - \mathcal{H}(q(\{\theta_d, \mathbf{z}_d\}))$$

$$\geq -\log p(\mathbf{W} \mid \alpha, \beta)$$

where l being the cost function, ξ being the slack in max-margin learning process. More specifically the max-margin constraints is expressed as:

$$\forall d, y, \log \frac{p(y_d \mid \mathbf{w}_d, \alpha, \beta)}{p(y \mid \mathbf{w}_d, \alpha, \beta)} \geq \Delta l_d(y) - \xi_d$$

Such optimization can be solved by applying linear expectation operator which was proposed in and applied in [8] and [7]. Then the variational algorithm proved in Corollary in [1], can be used to optimize over $q(\{\theta_d, \mathbf{z}_d\})$ and $q(\eta)$ and finally achieve the MedLDA model.

3. EXPERIMENTS AND METHODS

The experiments is designed to exploit linguistic contents of Tweets to obtain an efficient topical representation and to discover the undering user distribution. The interesting MedLDA modeling process is implemented and provided as open source package on GitHub [9], which can be configured and run on a distributed cluster on cloud.

3.1 Challenges and Limitations

One of the major challenges of this project is to decide how and where to collect Twitter data, and what to collect. Fortunately Twitter API provides convenience for developers who need to crawl Twitter data. However due to the request rate limits imposed by Twitter, collecting a large set of data is extremely time consuming. Choosing what to collect is also challenging. Since user can post in different languages and in heterogeneous characters and emoticons, abbreviation and concatenation of long phrases, or abuse of punctuation as opposed to classical alphabets and legitimate English words. Another challenge is to configure the cloud environment for MedLDA modeling. Amazon EC2 clusters can be set up with multiple virtual Linux machines that run in parallel and communicates through SSH connection, so as to speed up training and testing purposes.

Table 1: Political Orientation Prediction Accuracy

Machine 0	Machine 1	Machine2
0.799087 ± 10^{-3}	0.803655 ± 10^{-3}	0.753668 ± 10^{-3}

3.2 Experiment Approaches

Qualified users for each pre-defined label as introduced in section 1 are scripted from Twitter API using user search requests and stored in files accordingly to avoid excessive requests. Then another script would requests the timeline of each user iteratively and construct the training and testing data sets in the required format. Around 1000 users are selected for each label in both sets (8 labels for political orientation and 4 labels for interests), with 30% of them randomly chosen to be the training set, 30% to be cross-validation set, while the rest written into testing file. A list of words to ignore is used to avoid scripting out words that are only useful for constructing correct grammar: e.g., ”at”, ”of”, ”have”, ”are”,etc. Other neglected information from user timeline is the Twitter accounts mentioned by ”@”, and URLs or multimedia contents as they are not relevant to constructing topical representation for users. Hashtags are retained as they contain important information about what topics the user cares about. Constructed data files are then send to the cluster, where it will be evenly split into three sub files for each machine to process.

3.3 Analysis and Discussion

3.3.1 Political Orientation

With cross-validation between training and validation set of users, the number of topics is chosen to be 30 for the prediction purpose. With 3 testing machine working in parallel, each machine produces the following testing accuracy in Table 1:

To facilitate the observation of prediction datasets,

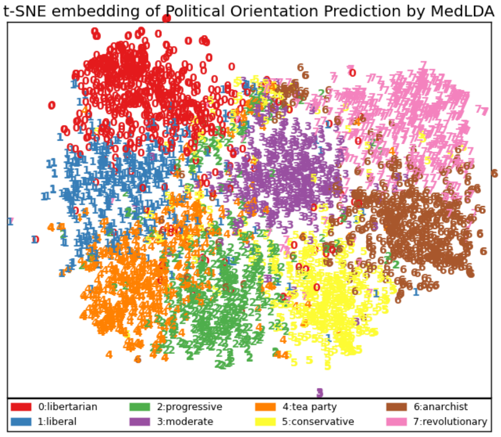

Figure 1: t-SNE Embedding of Political Orientation Prediction Result by MedLDA

a script that performs dimensional reduction is implemented using Skit-learn machine learning package in Python. The following 2-dimensional plotting reflect the label assignment and proximity of user-political orientation distribution. The Distributed Stochastic Neighbor Embedding model is used to visualize the result as in Figure 1:

Each color represents one political orientation label that is also distinctively associated with a number from 0 to 7, and each user that is predicted to be holding the orientation is plotted by the number with the associated color. It can be observed that most of the users can be correctly classified, while each label cluster also infiltrate one another, as users tend to take a middle ground in certain issues, or engaged in discussions or arguments on multiple sides of opinions. As also shown in Table 3, the top 10 most frequently used words by each political orientation correlates to one another, some share relatively high similarity.

3.3.2 Interests

Prediction accuracy and embedding results on personal interests are also obtained and visualized in Table 2 and Figure 2 accordingly in the same approach as political orientation set. As can be expected, the four types of interests (technology, music, sports and gaming) are of heterogeneous nature and therefore low correlation in corpus distribution. Thus the colored and labeled clusters in Figure 2 has rather low overlapping compared to the political data sets. The users with interests in technology has more overlapping with users with video gaming interests than with the other two groups of users, whose correlation can also be shown in top frequent words in Table 4.

Table 2: Interests Prediction Accuracy

Machine 0	Machine 1	Machine2
0.911990 ± 10^{-3}	0.909554 ± 10^{-3}	0.915816 ± 10^{-3}

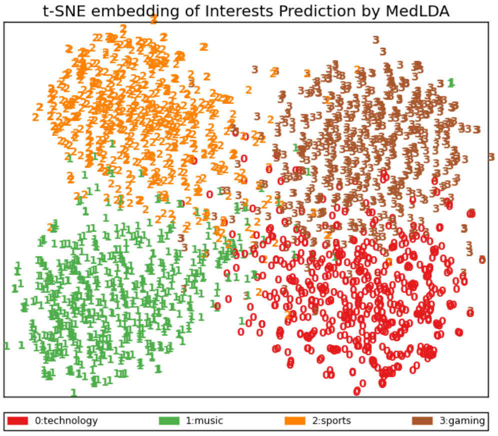

Figure 2: t-SNE Embedding of Interests Prediction Result by MedLDA

4. CONCLUSION AND FURTHER WORK

It can be concluded from the experiments above that MedLDA is a efficient topic model that can be used to learn topical representation and a reliable distribution of classifiers for prediction. Binary classification tasks that applied generic LDA on political orientation (democrat / republic) achieved accuracy as shown in [5] Table 2: LING-GLDA, which is lower than MedLDA achieved on average. It shows that multi-class classification of Twitter users on a specific aspect is possible and results can be accurate.

The experiment results shows possibility for further improvement on the online social network users attribute classification and prediction. Options can be incorporating users' network connection information to study the transfer and spreading of information, such that to predict user behaviors before a new viral marketing plan is rolled out. Linguistic sentiment analysis can also be conducted alongside with user classification so as to obtain more accurate prediction not just on whether user would be interested in upcoming social network trend, but also to predict what point of standing the user would be likely to take.

5. REFERENCES

[1] Zhu, Jun, Amr Ahmed, and Eric P. Xing.

MedLDA: maximum margin supervised topic models. The Journal of Machine Learning Research 13, no. 1 (2012): 2237-2278.
[2] David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, (3):9931022, 2003.
[3] Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In Advances in Neural Information Processing Systems (NIPS), pages 470476, Denver, Colorado, 1999.
[4] Tony Jebara. Discriminative, Generative and Imitative Learning. PhD thesis, Media Laboratory, MIT, Dec 2001.
[5] Pennacchiotti, Marco, and Ana-Maria Popescu. A Machine Learning Approach to Twitter User Classification. ICWSM 11 (2011): 281-288.
[6] Rao, D.; D., Y.; Shreevats, A.; and Gupta, M. 2010. Classifying Latent User Attributes in Twitter. In Proceedings of SMUC-10, 710718.
[7] Jun Zhu and Eric P. Xing. Conditional topic random fields. In J. Fu rnkranz and T. Joachims, editors, International Conference on Machine Learning (ICML), pages 12391246, Haifa, Israel, 2010. [8] Jun Zhu, Eric P. Xing, and Bo Zhang. Partially observed maximum entropy discrimination Markov networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems (NIPS), pages 19771984, 2008. [9] ptuum package from SailingLab at Carneige Mellon University: https://github.com/petuum
And Tables for most frequent words in each topic:

Libertarian	Liberal	Progressive	Moderate	TeaParty	Conservative	Anachist	Revolutionary
#tlot	new	vote	politicians	teaparty	#tcot	anarchist	revolution
libertarian	people	GOP	GOP	#tcot	Obama	anarchism	people
#tcot	Obama	lgbt	Obama	Obama	baltimore	#dumbsheeple	must
#ronpaul	community	gay	#morningjoe	conservative	conservative	police	#marx
paul	GOP	Obama	#TPP	Palin	Hillary	#bahrain	revolutionary
free	baltimore	baltimore	#tcot	GOP	#uniteblue	protest	capitalist
people	#auspol	people	Cliton	politics	Clinton	people	class
government	#libdems	state	#nofasttrack	Hillary	teaparty	solidarity	American
state	#uniteblue	support	bill	#Mesa	GOP	anti	#tcot
liberty	liberal	#coleg	police	people	police	revolution	evil

Table 3: Top 10 most frequent words for each political orientation

technology	sports	music	gaming
Google	game	music	game
Apple	live	fashion	review
#cdntech	sports	UK	ps4
Facebook	#NFL	album	#IGN
Tech	#nfldraft	artist	xbox
blog	time	soundcloud	#gameinsight
data	win	video	iphone
Twitter	team	free	live
mobile	league	musician	playstation
app	season	show	blog

Table 4: Top 10 most frequent words for each interest

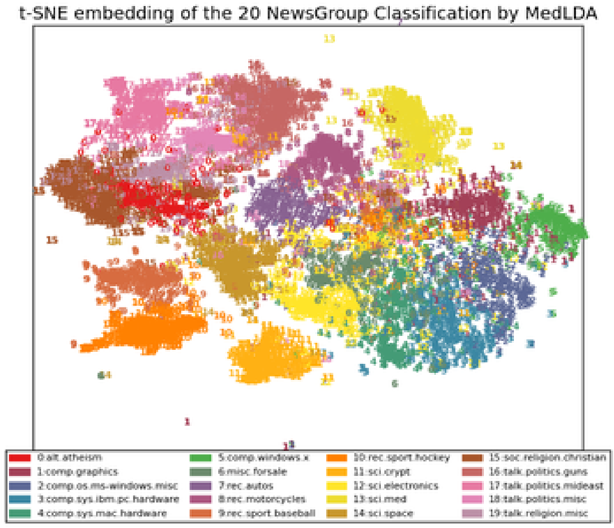

Figure 3: t-SNE Embedding of 20 News-Group Prediction Result by MedLDA (open dataset for testing the cloud cluster and embed plot script, result can be compared to similar experiment in [1])