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 (LDA), 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[7].

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.

3. EXPERIMENTS AND METHODS

The experiments is designed to exploit linguistic content of Tweets only to discover

3.1 Challenges and Limitations

One of the major challenges of this project is to decide how and where to collect data. Fortunately Twitter API provides convenience for developers who need to crawl Twitter data.

3.2 Experiment Approaches

3.3 Analysis and Discussion

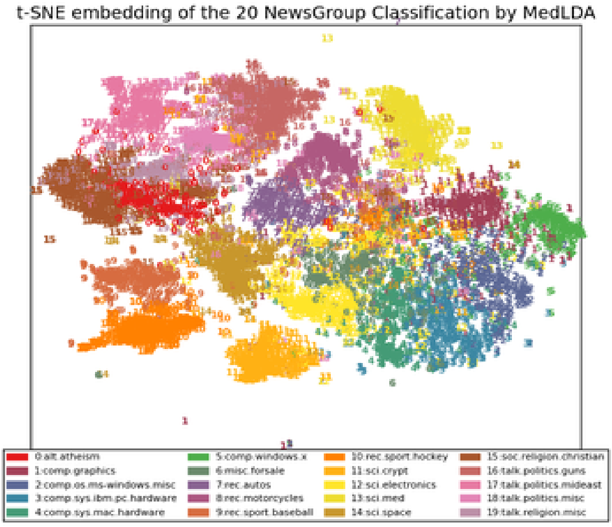

Figure 1: t-SNE Embedding of 20 NewsGroup Classification Result by MedLDA

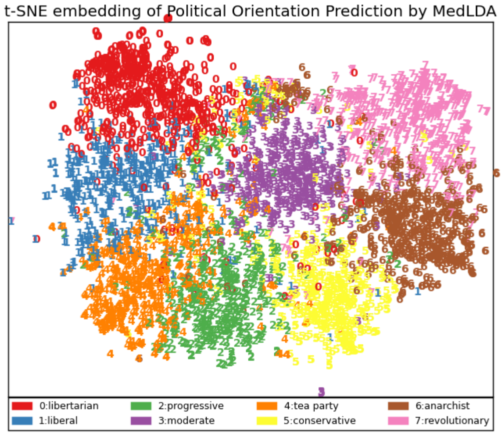

Figure 2: t-SNE Embedding of Political Orientation Prediction Result by MedLDA

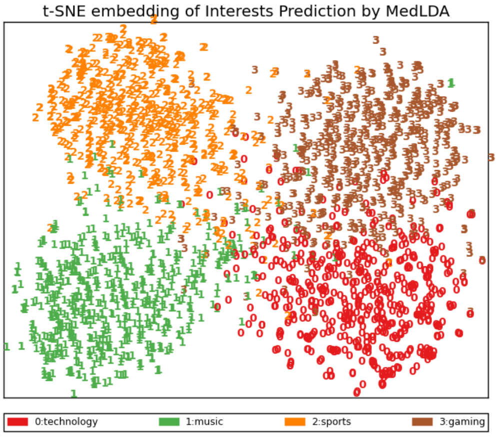

Figure 3: t-SNE Embedding of Interests Prediction Result by MedLDA

4. CONCLUSION AND FURTHER WORK

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] 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	TeaPa
#tlot	new	vote	politicians	teapa
libertarian	people	GOP	GOP	#tco
#tcot	Obama	lgbt	Obama	Obam
#ronpaul	community	gay	#morningjoe	conserv
paul	GOP	Obama	#TPP	Pali
free	baltimore	baltimore	#tcot	GO
people	#auspol	people	Cliton	polit
government	#libdems	state	#nofasttrack	Hilla
state	#uniteblue	support	bill	#Me
liberty	liberal	#coleg	police	peop

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