A SENTIMENT-TOPIC MODEL FOR TWITTER

YUNFEI LU AND NIKITA NANGIA

ABSTRACT. Politicians, reporters, and the general populous rely heavily on polls and predictions made by various experts for the prediction of electoral results. The failure of recent predictions made by such experts in the United Kingdom and the United States motivates the search for better tools and techniques. This objective of this paper is to explore new avenues of topic modeling for twitter. The main contribution made is using labeled sentiment data in a LDA model, in the style of an author-topic LDA model (Rosen-Zvi et al., 2004)

1. INTRODUCTION

With the advent of social media platforms like Twitter and Facebook, short texts in the form of updates and tweets have become an instrumental source of information. Even though each tweet is no more than a 140 characters, the cumulative sum of information available on Twitter is astonishing and complex. To be able to efficiently interpret such short texts then is vital.

Aside from being a source of news, Twitter is largely a source of public sentiment information. Considerable research has been devoted to sentiment and opinion analysis of the Twitter corpus (Go et al., 2009; Pak et al., 2010; Agarwal et al. 2011).

2. TOPIC MODELING

Topic modeling has been researched for years and is still gaining increasing attention. Latent Dirichlet Allocation is one of the standard methods and has been extended or particularized in a variety of ways. For example, Wang et al. proposed a topic model to analyze image corpora in order to solve computer vision problems. Blei et al. introduced a supervised LDA model for better discriminant analysis. Rosen-Zvi et al. offered an author-topic model, which can model authors and their corresponding topic distributions. In their paper, they found that the model shows better performance than standard LDA when only a small number of words can be obtained from the documents. Applying topic models for short or few documents for text clustering is more challenging because of data sparsity and the limited contexts in such texts. One approach is to assume there is only one topic

per document(Nigam et al., 2000). This method is called Dirichlet Multinomial Mixture(DMM) model. In this model, each document is assumed to have only one topic. The process of generating a document d in the collection of documents D, as shown in Figure 1, is to first select a topic for the document, and then all the words in the documents are generated based on the topic to word Dirichlet-multinomial component. Biterm-Topic Model is a word co-occurrence based topic model that learns the topic by modeling word-word co-occurring in the same context. For example, in the same short text window. Unlike LDA models the word occurrences, BTM models the biterm occurrences in a corpus. In generation procedure, a biterm is generated by drawn two words independently from a same topic. In other words, the distribution of a biterm b = (wi, wj) is defined as: p(b) =

3. INTUITION

The author-topic model has been proved is better than LDA when dealing with the short text data. The main reason,

The work has been done in this project is: (1) Several algorithms are tested and results are given. (2) Sentiment-Topic model is implemented based on the Author topic model.

4. DATA SELECTION

In this experiment, twitter streaming APIs is used to get the data. The streaming APIs give developers low latency access to Twitter's global stream of Tweet data. The filter is used to only get the data in the Great New York area and only English is accepted at this time. The timestamp of the data is December 15th. The raw data is JSON format and a simple parser is written in order to get the valid text data. Since some users would retweet other's tweet and simply comment few words. In order to make this kind of data meaningful, the current tweet will be combined with the tweet which is retweeted from together as a valid data. All tweets will be combined with their hashtag to become valid data. NLTK package is used to remove the stop words and Stemmers is used to remove morphological affixes from words, leaving only the word stem. All punctuations are removed. Emojis and URLs are also removed. Since all algorithms are used in this paper is based on the bag-of-words, all sentence are parsed as a number of tokens and save in the file.

4.

REFERENCES

[1] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smith, P. 2004. Theauthor-topic model for authors and documents. In *Proceedings of*

the 20th Conference on Uncertainty in Artificial Intelligence, pages 487-494. AUAI Press.

[2] Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of LREC*.

[3] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Technical report*, Stanford.

[4] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. In *Proceedings of the Workshop on Language in Social Media*, pages 30-38.