Topic mining for short-text documents on micro-blogging site by combining Doc2Vec and clustering techniques

Thuong-Hai Pham
Faculty of Information and Communication Technology
University of Malta
Msida MSD 2080, Malta
Email: thuong-hai.pham.16@um.edu.mt

***Abstract*—In the era of data explosion these days, especially digital text generated by World-Wide-Web users, it demands techniques that automatically organize large collection of text for further analysing and processing tasks. One family of those techniques is called "topic model". These techniques discover underlying topics from a given corpus with or without intervention of human, in other word, supervised and unsupervised. This report is aimed to examine the traditional Latent Dirichlet Allocation (LDA) and a proposed method of combining Doc2Vec and clustering technique on the problem of topic mining. For practical evaluation, Twitter[1] is chosen to do experiment on three approaches: standard LDA, author-topic model and our proposed approach. The paper also covers background behind each method and proposes a plan to conduct this research on these approaches with estimated time-line and resources.**

1. Introduction

Applying topic model for micro-blogging site is a very important task to enhance our understanding of the social network. One very successful technique and also being considered as state-of-the-art in unsupervised topic model is Latent Dirichlet Allocation (LDA) [1]. Some applications were proposed by Zhao et al. (2011) [2] with the work of comparing Twitter and traditional media by LDA, or finding topic-sensitive influencers on Twitter by Weng et al. (2010) [3].

It is important to be noted that applying LDA directly on micro-blogging sites is considered to be not a trivial task yet more challenging problem. This occurs due to the nature of micro-blogging sites which is the limited length of each posting unit (e.g. tweets on Twitter have maximum 140 characters each). In addition, proposed LDA solution while trying to solve this problem have to make more assumptions (single-topic tweets...) [2] about the data itself rather than the bag-of-words (BoW) assumption from original LDA.

Therefore, we consider examining clustering method such as K-means to discover the underlying topic. The feature learning is done by doc2vec of Le & Mikolov [4], which is an adaptation of word2vec [5].

1. https://twitter.com/

For the rest of this paper, in section 2, we discuss about the mathematical background behind BoW assumption for topic model: the infinite exchangeability and De Finetti theorem. Thereafter, we revise LDA as a generative probabilistic model and its latent variables in section 2.2. However, LDA does not work well when applying directly to short-text documents as tweets, we then consider two variants of LDA to solve this problem which are author-topic model (2.3.1) and Twitter-LDA (2.3.2). To end with the background and related works, the three groups of methods to evaluate topic models are also mentioned in section 2.4. To the most important part, we figure out the disadvantages of these methods and present our proposal (section 3). Also in this section, we breakdown the required tasks, its timeline, and needed resources with expected result and challenges we may face.

2. Background

2.1. Mathematical background

2.1.1. Exchangeability. We say that $(x_1, x_2...)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $p(x_1, x_2, ..., x_n)$ is invariant to permutation of the indices. That is, for any permutation π,

$$p(x_1, ..., x_n) = p(x_\pi(1), ..., x_\pi(n))$$

It is important to emphasize that independent and identically distributed random variables are always infinitely exchangeable. However, infinite exchangeability is a much broader concept than being independent and identically distributed.

2.1.2. De Finetti theorem, 1935. A sequence of random variables $(x_1, x_2, ...)$ is infinitely exchangeable iff,for all n,

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta)$$

for some measure P on θ. If one assumes the data is infinitely exchangeable, then there must exist an underlying parameter and prior.

2.2. Latent Dirichlet Allocation

LDA is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where a topic is characterized by a distribution over words. To implement this idea, LDA assumes each document is a bag of words (BoW assumption). Hence, it applies infinite exchangeability on the documents and inherits the De Finetti theorem to expect an latent parameter and prior underlying in the corpus. These latent variables are illustrated in the figure 1 below.

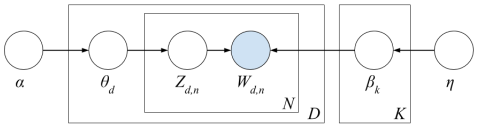

Figure 1. LDA graphical model

In figure 1:

- α is Dirichlet distribution parameter, controls the shape and sparsity of θ
- θ are per-document topic proportions.
  θ is a K-dimensional Dirichlet random variable, takes values in the (k-1)-simplex, and has the following probabiltiy density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_K^{\alpha_K - 1}$$

  The Dirichlet is conjugate to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- $Z_{d,n}$ is per-word topic assignment, in which D and N are number of documents and number of words in a specific document, respectively.
- $W_{d,n}$ is observed word.
- β are topics, which is V dimensional Dirichlet.
- η is the topic hyper parameter.

The blue-shaded node denotes observed variable, the others are hidden or latent variables. Plates denote replicated structures.

From a collection of documents, LDA infers: per-word topic assignment $Z_{d,n}$, per-document topic proportions θ_d and per-corpus topic distributions β_k.

2.3. Latent Dirichlet Allocation variants for Twitter

One very basic approach is to apply LDA directly to Twitter by treating each tweet as a single document. However, due to the constraint of 140 characters per tweet, a tweet is too short for LDA to figure out the topic proportions.

2.3.1. The author-topic model. To overcome this issue, by excluding the topic proportions for each tweets but taking into consideration only the underlying topics in each user, aggregating all tweets of a Twitter's user into a single document was proposed and gained a better result to direct LDA [3], [6].

On one hand, this approach is very efficient on a specific task (e.g. topic-sensitive influencers mining [3]) by not modifying the inference process of original LDA. On the other hand, the target of this approach is each user, not tweets, so it is not applicable for a general problem of topic mining.

2.3.2. Twitter-LDA. On a different perspective, while attempting to compare Twitter and traditional media, Zhao et al. [2] proposed Twitter-LDA, a modified version of LDA to work on Twitter's short tweets without concatenating all tweets into one.

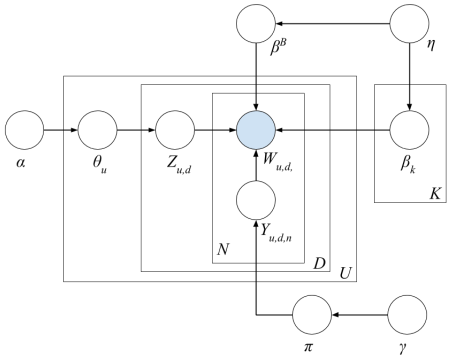

Figure 2. Twitter-LDA graphical model

In figure 2, the author introduced four more variables:

- β^B denotes the background words distribution
- π denotes a Bernoulli distribution which simulate the choice of authors between topic-related words and background words.
- γ is the parameter of distribution π.
- $Y_{u,d,v}$ denotes the selection of background or topic word.

and a slightly modification on θ that θ_u represents per-user topic proportions instead of per-document as in the original version. In addition, the D (document) plate is surrounded by a new plate U which stands for each user. It is necessary to mention that Twitter-LDA makes an assumption in which each tweet only conveys a single topic. We will clarify our disagreement on this assumption in section 3.

2.4. Evaluation

Wallach et al. [7] summarized a variety of methods to evaluate LDA. As a topic model method, LDA is commonly evaluated by intrinsic and extrinsic evaluation.

2.4.1. Intrinsic evaluation. One very basic intrinsic evaluation method is to view the problem as document modelling [1]. By stating that, the goal of the model is to achieve high likelihood on a held-out test set, C'. In this case, the

perplexity measure is used as in normal language modelling problem, in which the lower perplexity, the better performance the model achieves.

$$perp(C') = exp\left\{-\frac{\sum_{d=1}^{D} log(p(W_d))}{\sum_{d=1}^{D} N_d}\right\}$$

More advanced, measurement is also estimated by the probability of unseen held-out documents given some training documents. This probability can be written as [7]

$$P(C|C') = \int d\theta d\alpha dm P(C|\theta, \alpha m) P(\theta, \alpha m|C').$$

in which, C, C' denote training documents set (corpus) and held-out documents set, respectively. Noted that m is the base measure of Dirichlet distribution, in addition to the concentration parameter α.

2.4.2. Extrinsic evaluation. On the other hand, extrinsic approaches measure LDA performance on some secondary tasks, such as corpus comparison [2] or topic-sensitive influencers mining [3]. These approaches are similar to the way language models performance are measured.

2.4.3. Human evaluation. As a part of the corpus comparing work [2], Zhao et al. also evaluated the performance between original LDA, author-topic model and their proposed Twitter-LDA. Based on preliminary experiments, the authors set number of topic K to 110 for each model, then mixed 330 topics from the three models. The topics were then scored by two human judges. Each assigned a score on each topic, ranging from 1 (meaningful) to 0 (nonsense).

3. Proposed method

The two methods of author-topic model and Twitter-LDA have overcome the problem of short tweets in the micro-blogging site Twitter which prevent direct usage of original LDA. More than that, Twitter-LDA has showed to outperform the other two in experiment. Nevertheless, Twitter-LDA has to change the original LDA process and inference approximation algorithm to implement its idea. This approach is hard to be re-implemented in industrial sector by using existed library for other problems.

More than that, it is worth to consider the assumption of one tweet belongs only to one topic. One topic may represents only one topic intended by its Twitter user. However, please note that this topic is not the final topic discovered by our mining methods but may be a combination of the two final topics. For example, the topic user intend to tweet about is public insurance. Throughout the whole corpus, our mining process points out two distinct topics: healthcare and public administration. It is obvious that the user-intended topic reflects both of our discovered topics on the perspective of the whole corpus.

Bearing that in mind, we would like to propose a method to compare with original LDA and author-topic model, by accepting only BoW assumption.

3.1. Clustering based on distributed representation of sentences

3.1.1. Distributed representation of sentences for features learning. After word2vec [5], Le and Mikolov presented the paragraph (document) vector models. Formally, the objective of a word embedding model is to maximise the log probability

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k}, \ldots, w_{t+k})$$

given a sequence of training words $w_1, w_2, w_3, \ldots, w_T$.

The paragraph vector models uses the same idea to develop the paragraph vector framework. In fact, this is not a single model but implemented into two different approaches: Distributed Memory model (DM) and Distributed Bag of Words (DBOW) - vector without word ordering.

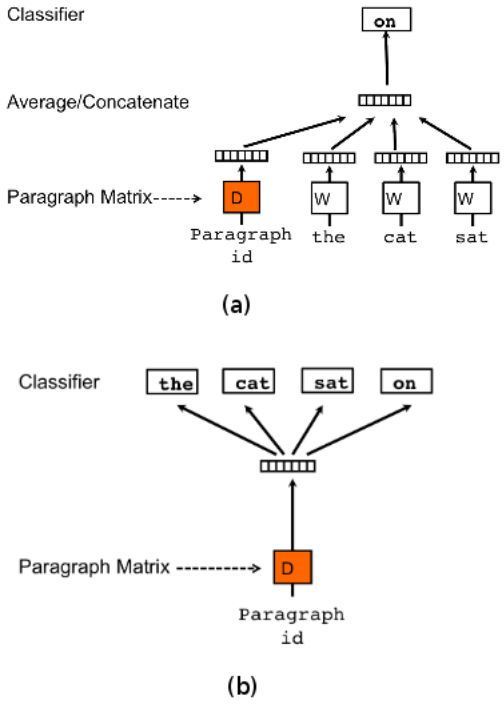

Figure 3. Framework for learning paragraph vector, (a) Distributed Memory, (b) Distributed Bag of Words

In figure 3 above[2], the DM model (a) is actually constructed from each word vector of that structure (sentence/document), then these vectors are combined (through averaging or concatenation) to learn the sentence/document features. In addition, a paragraph matrix is maintained to keep track of the whole sentence/document. In contrary to that, the DBOW model does not combine any word vectors yet only a paragraph vector is trained to predict the context.

2. https://arxiv.org/pdf/1405.4053v2.pdf

WBS	Task description	Start date	Finish date	Progress
1	**Data acquisition**	**1/30/2017**	**2/1/2017**	**75%**
1.1	Extract data from Twitter archive	1/30/2017	2/1/2017	100%
1.2	Cleaning tweets data	2/1/2017	2/1/2017	0%
2	**Implementation**	**2/2/2017**	**2/9/2017**	**0%**
2.1	Develop code for LDA	2/2/2017	2/4/2017	0%
2.2	Develop code for author-topic model	2/6/2017	2/6/2017	0%
2.3	Develop code for clustering: doc2vec & K-means	2/7/2017	2/9/2017	0%
3	**Experiments**	**2/13/2017**	**2/21/2017**	**0%**
3.1	LDA	2/13/2017	2/21/2017	0%
3.2	LDA with author-topic	2/16/2017	2/21/2017	0%
3.3	Clustering	2/20/2017	2/21/2017	0%
3.4	Clustering with author-topic	2/20/2017	2/21/2017	0%
4	**Evaluation**	**2/22/2017**	**2/23/2017**	**0%**
4.1	1st judge topic assignment	2/22/2017	2/22/2017	0%
4.2	2nd judge topic assignment	2/22/2017	2/22/2017	0%
4.3	Mutual agreement scores	2/23/2017	2/23/2017	0%
5	Final Report	2/24/2017	2/25/2017	0%

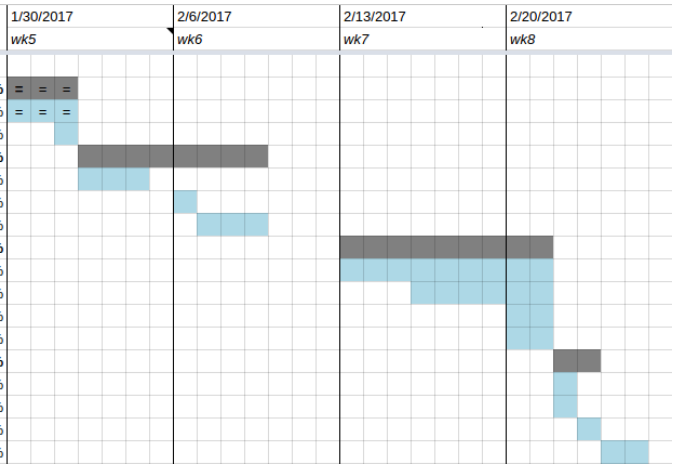

Figure 4. Project timeline

3.1.2. Clustering based on the learned features. Our proposed method consists of two distinguished phases. First, "meaning" or, in fact, features of each document is learned by doc2vec model as presented above. Having the features, the documents are then clustered using clustering technique.

For the goal of our experiment described latter in 2.4, we choose K-means to find hard clusters from our documents, each cluster represents a topic. However, to produce a probabilistic topic proportions as in LDA, we can easily change K-means to C-mean to achieve fuzzy or soft clusters.

3.2. Tasks

The project is planned into four main phases: data acquisition, implementation, experiments and evaluation phase. Figure 4 illustrates sub-tasks in each phases and their estimated required working time.

To be more detailed, data for this project is acquired from Twitter stream archive[3] in July 2017. Although we do not need to stream data directly from streaming API, data preprocessing still plays an essential role to filter out other languages (Chinese, Japanese, Spanish...) tweets, remove stop-words, tokenize words within tweets, remove urls and normalise unconventional language used on social media (character repetition, emoticons...)

Source-code used for the latter evaluation is developed using LDA and Doc2Vec model in Gensim[4] library.

For the evaluation task, although perplexity is considered to justify the result with less subjectivity, it does not measure how meaningful the topics discovered. Hence, we make use of the human evaluation strategy [2] instead. This evaluation process is performed by two distinct judges. Each judge assigns name for all of the topics in our output results. Afterwards, their topic name assignments are exchanged to each other to score from 0 to 10 how they agree with the other judge about the topic names. The meaningfulness of each algorithm and parameter pair is measured by averaging all the agreement score above, which basically reflect the interpretability of the result.

3.3. Resources

Due to the magnitude of our data (48.7GB in compressed format), a sufficiently efficient machine is required to perform data preprocessing and calculation for our experiment. Therefore, we intend to make use of the Google cloud compute engine n1-standard-4[5]. In addition, the experiment evaluation process also required to employ 2 judges (the more, the better) to individually score meaningfulness of our topic mining result.

4. Expected result & difficulties

Our hypothesis for this is the proposed method of combining Doc2Vec and K-means achieves approximate meaningfulness score in compare to traditional LDA method. This expectation also implies that the proposed method can perform equally or better by having no assumption, which is subjective and arguably untrue, about the data itself.

On the other hand, we anticipate to face with some difficulty regards to data. Firstly, Twitter stream archive is too large and prevent us to consume all of the available tweets for our models. Hence, we have to prioritise tweets to be fed to our models by the number of tweets per user. Only top tweeted users are being taken into consideration. This strategy also support in making the difference when running author-topic model. Secondly, the acquired data may contains too much noise and unconventional language usage. Thirdly, the meaning of topic discovered can be unclear or hardly interpretable because of the fact that most tweets are about their daily life instead of broad topic coverage in newspaper. These two latter obstacles can be solved in some degree by careful cleaning of raw data in our data acquisition phases.

3. https://archive.org/details/archiveteam-twitter-stream-2016-07
4. https://radimrehurek.com/gensim/
5. operates with 4 virtual CPUs, 15GB RAM, 200GB hard disk drive

References

[1] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European Conference on Information Retrieval*, Springer. Springer Berlin Heidelberg, 2011, pp. 338–349. [Online]. Available: http://ink.library.smu.edu.sg/sis{_}researchhttp://ink.library.smu.edu.sg/sis{_}research/1375http://link.springer.com/10.1007/978-3-642-20161-5{_}34

[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank," in *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, 2010, p. 261. [Online]. Available: http://ink.library.smu.edu.sg/sis{_}researchhttp://portal.acm.org/citation.cfm?doid=1718487.1718520http://ink.library.smu.edu.sg/sis{\{}{_}{\}}research

[4] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[6] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, ACM. New York, New York, USA: ACM Press, 2010, pp. 80–88. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1964858.1964870

[7] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. New York, New York, USA: ACM Press, 2009, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1553374.1553515