Topic modeling for social media content: A practical approach

Article · August 2016

CITATION
1

READS
511

2 authors, including:

Ghazaleh Babanejad
York University
10 PUBLICATIONS **32** CITATIONS

SEE PROFILE

Topic Modeling for Social Media Content: A Practical Approach

Vala Ali Rohani
Department of Data Analytics
Berkshire Media
Kuala Lumpur, Malaysia
vala@BerkshireMedia.com.my

Shahid Shayaa
Berkshire Media
Kuala Lumpur, Malaysia
shahid@BerkshireMedia.com.my

Ghazaleh Babanejaddehaki
Faculty of Computer Science and Information Technology
Univerity Purtra Malaysia
Kuala Lumpur, Malaysia
gs377178@student.upm.edu.my

***Abstract*— Perceiving the discussed topic in social media brings a great amount of value to different fields, such as marketing, security, education, and management. Topic modeling provides a powerful method for projecting text documents into topic space. In this paper, we explore an unsupervised topic modeling approach that incorporates LDA algorithm toward discovering the topics in social media content. Empirical experiments on social media datasets with 90,527 records reveal that this approach is quite effective for detecting the topic facets and extracting their dynamics over time. The studied model is quite general and can be applied in a wide variety of domains to automatically mining topics from the social media channels.**

Keywords—Topic Modeling, Latent Dirichlet Allocation, Text Mining, Social Media

I. INTRODUCTION

By emergence of social media, people became more curious to express their opinions on web about their day-to-day activities along with global events. Evolution of Web 2.0 has also contributed extremely to these activities, thereby providing us an easy to use platform to share very huge amount of structured and unstructured data across the world, referred to as Big Data [1]. The huge number of social networks, blogs, forums, news reports, online marketplaces, and additional web resources serve as platforms to publish contents, which can be utilized for understanding the opinions of the general public and consumers on social events, political movements, company strategies, marketing campaigns, product preferences, and also monitoring reputations. Such data can be analyzed using a combination of Data Mining, Web Mining and Text Mining techniques in a wide range of real life applications [2, 3].

Understanding large collections of unstructured content with high level of accuracy is still a persistent challenge in data mining [4]. Unlike information retrieval, where users know what they are looking for, in such cases, users are more interested in detecting the high-level themes of textual contents, called topics. Addressing this desire, Topic models were emerged as effective tools to discover latent text patterns in the content [5]. Accordingly, Topic modeling became a popular method to elicit thematic structure from large amount of textual contents without human supervision [6]. It is a specific branch of text mining which identifies patterns in a corpus. To this end, the related words are grouped into topics which are defined as a recurring pattern of co-occurring words.

One of the most familiar methods in topic modeling is Latent Dirichlet Allocation (LDA). As a generative probabilistic model for collections of discrete data, LDA is based on a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. In turn, Each topic is considered as an infinite mixture over an underlying set of topic probabilities [7]. Topic models such as the LDA have become a ubiquitous and effective tool in machine learning [8]. They have played a significant role in a variety of data mining tasks, within a wide scope of research activities, including computer science [7, 9, 10], biomedical informatics [11], scientometrics [12], social and political science [13-15], and digital humanities [16].

The goal of our research is to develop a practical topic model based on LDA to elicit topics from social media corpus. We studied social media datasets with 90,527 records in domain of aviation and airport management. To improve the accuracy of implemented topic modeling algorithm, we identified a list of generic keywords including 645 English and Malay common words which were excluded from the studied datasets. In our proposed method, the model accurately detected five main topics discussed in social media in Nov 2015 within the studied domain of aviation. And furthermore, the dynamics of topics was visualized per day based on a probabilistic model.

The rest of the paper is organized as follows. Section 2 discusses some previous approaches on topic modeling and its applications. Section 3 explains the applied topic modeling method called Latent Dirichlet Allocation (LDA). Data sampling, cleaning and analysis of social media content are presented in Section 4. Finally, Section 5 concludes with a discussion of how topic modeling can be utilized to address real world problems.

II. Related Works

Recent studies in a variety of research areas show increasing interests in topic modeling. The early research on topic modeling conducted by Deerwester et al. in 1990 which proposed a semantic structure for improving detection of relevant documents based on terms found in queries [17]. Papadimitriou et al. in 1998 presented the technique of random projection as a way of speeding up LSI method [18]. Another research was done by Hofmann in 1999 which proposed a new model of Probabilistic Latent Semantic Indexing (PLSI) to deal with domain specific synonymy as well as with polysemous words [19]. Later in 2003 the Latent Dirichlet allocation (LDA) developed by David Blei et al. they worked on allowing documents to have a mixture of topics [7]. LDA became a standard tool in topic modeling. There are so many authors working on LDA. They have extended it in different ways especially in social networks and social media. In 2008, Beli and his friend Mcauliffe proposed the improved LDA as supervised LDA (sLDA) which was a statistical model of labeled documents. They infer a most likelihood procedure for parameter estimation [20]. Another interesting model introduced by McCallum in 2007 which concurrently found groups between entitles, and topics between corresponding text [21]. Also in 2007 Zhang et al. worked on hierarchical Bayesian model inferred from the widely-received LDA model. They found probabilistic community profiles in social networks. In this model, communities are considered as latent variables and defined as distributions over the social actor space[22]. Later in 2008 Nallapati et al. did the same thing. They produced a model to combine LDA with community-detection process [23].

In 2009 Chang et.al worked on the probabilistic topic model for analyzing text structure and derived the related title to them. After that, they found the relationship between them and Wikipedia [24]. In 2009 Liu et al. developed a Bayesian hierarchical approach which it could execute topic modeling and author community discovery in one combined framework [25]. Hong et al. studied on the problem of using standard topic models in micro-blogging environments in 2010. They proved that by training a topic model on accumulated messages the quality of learned model will increase and the performance will be significant in real world classification [5]. Another interesting research in this domain was done by Li et al. in 2010. They combined LDA and Girvan-Newman community detection algorithm and proposed TTR-LDA. This algorithm was applied over different datasets. The results showed users in the same community want to be interested in a similar set of topics. Also it was concluded that topics could be divided in to sub-topics and spread into various communities over time [26]. In 2011, Wang and Beli combined traditional collaborative filtering with probabilistic topic modeling. This algorithm recommended scientific articles to users of an online community. This recommendation covers all newly and published articles [10]. Zhai et al. in 2012 introduced a flexible large scale topic modeling package in MapReduce. This model could find topics and extract them from different language [27]. Daud in his research in 2012 presented a topic modeling algorithm named it Temporal-Author-Topic (TAT). This model could simulate model text, researchers and the time that research was done. This algorithm used semantic-based intrinsic of words between authors and papers. This model could find related researchers in a different period of time in same topic research. Also, this model showed the change of author's relationships and interests over time [28].Vavliakis et al. in 2013 defined the concept of an important event and proposed a method for performing event detection from web documents in large time-stamped. They detect all important events from document stream [29]. In same duration Lim et al. in their framework introduced Twitter Network topic model which is worked in a full Bayesian nonparametric way. This model used hierarchical Poisson-Dirichlet processes (PDP) for text modeling and a Gaussian process random function model for social network modeling [30]. Also, in 2013 Nguyen and his friends discovered a model for supporting products marketing in social networks. They proposed a content-based social network analysis to find out the desired topics [31]. In 2014 Yang et al. presented a range of ideas for topic modeling techniques which contribute to deployed system. Also, they represented an algorithm for tweet text classification and close-loop for combining text with additional sources of information [32]. In one of the most recent research works in this domain in 2015, Ostrowski discovered new model which works based on classification as well as identification of noteworthy topics when it is applied to a filtered collection of twitter messages [33].

III. Topic Modeling – Latent Dirichlet Allocation (LDA)

LDA is a technique which automatically finds topics that contain certain words. Actually it represents mixture of topics that discharge words with certain probabilities. LDA can discover automatically the topics in a collection of documents. The goal of LDA is to derive the latent topic from words and documents. LDA first recreate the documents in the corpus. It works based on the importance of related topics from words and documents [7]. Assume that LDA has these specifications for generative process for each document w in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Now here there are some explanations:

1. The dimensionality k of Dirichlet distribution and the dimensionality of the topic variable as z are known and fixed.

2. The word probabilities are defined as $k \times V$ matrix β that $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, was considered it as a fixed quantity for estimation.
3. Poisson assumption is not important to anything that follows and the documents with more realistic length distributions can be used if needed.
4. N in independent from all data generating variables like θ and z. It is also a secondary variable and its randomness will ignore in the following development.

θ is a k-dimensional Dirichlet random variable that can take values in the (k-1)-simplex (a *k*-vector θ lies in the $(k-1)$-simplex if $\theta i \geq 0$, $\sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha i)}{\prod_{i=1}^{k} \Gamma(\alpha i)} \theta_1^{\alpha_1 - 1} \ldots \theta_\kappa^{\alpha_k - 1} , \quad (1)$$

α is a *k*-vector with components $\alpha i > 0$, and $\Gamma(x)$ is Gamma function. Given the parameters α and β, the joint distribution of a topic mixture θ, a set of *N* topics z, and a set of *N* words w is given by:

$$p(\theta, z, w|\alpha,\beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n |\theta) p(w_n |z_n, \beta) \quad (2)$$

where $p(zn |\theta)$ is simply θi for the unique *i* such that $z_n^i = 1$. Integrating over θ and summing over *z*, we obtain the marginal distribution of a document:

$$p(w|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n |\theta) p(wn |zn,\beta)\right) d\theta \quad (3)$$

Identifying LDA from a simple Dirichlet-multinomial clustering model is very important. The classical clustering model includes two-level model that a Dirichlet consider as a sample for a corpus once, a multinomial clustering variable also selected for each document in corpus. Also a set of words are selected for document conditional on the cluster variable. But LDA has three levels and the difference is that the topic node is sampled repeatedly during document. So one document can consider for different topics [7].

IV. Experiments and Results

In this section, we discuss our data sampling process and analysis of social media content toward discovering the related topics. As the dataset for this research work, 90,527 social media records in domain of Aviation and Airport Management in year 2015 was imported using Radian 6 software. After that, for every day of Nov 2015, a separate corpus was generated containing all social media content for that day, totally 30 text file as the input datasets for our experiments. In other words, all social media content of studied domain published in one day, was aggregated in a single dataset for that specific day.

To develop LDA topic modeling method, we used R [34] as a programming language and environment for statistical computing. Before running LDA algorithm over 30 input datasets, we cleaned the social media data in several steps. To this end, after removing punctuations, extra spaces, and other unnecessary patterns, we created a list of English and Malay stop words including over 600 common words to get deleted from the input datasets. This stage was quite effective in eliciting related words coherently for discovered topics. After that, LDA algorithm was applied on cleaned datasets to elicit the top social media topics for every day of studied domain.

For evaluating the performance of our developed model qualitatively, we asked domain experts to investigate the discovered topics along with assigned related keywords to select a title for each topic. Table 1 illustrates the list of top 5 topics detected by developed topic model for studied social media content. As shown in this table, the titles assigned to each topic by domain experts reveals that the proposed topic modeling algorithm correctly elicited related keywords for each topic.

TABLE I. Top 5 Discovered Topics in Studied Social Media Datasets

Topic 1 (KKIA)	Topic 2 (Paris Attack)	Topic 3 (Tony & Flood)	Topic 4 (CSR)	Topic 5 (Korean food)
terminal	security	tonyfernandes	csr	share
kkia	third	hmm	beyondborders	dkt
international	minister	flood	campaign	throwback
mahb	deputy	wet	beyond	lapar
airasia	paris	banjir	borders	avenue
disember	attacks	tony	english	korean
bki	government	fernandes	story	streetcafe
operations	terminal	501awani	englishday	denabahrin
operasi	aziz	kerajaan	drama	foodfor
news	tonyfernandes	dear	golden	cafe

Taking a quick look at Table 1 make it clear that the algorithm has done a pretty decent job. Topic 1, is about moving AirAsia flights from Terminal 2 to Terminal 1 at Kota Kinabalu International Airport (KKIA). Security issues in airports and also Paris terrorism attacks are categorized in Topic 2. Heavy rain in Nov 2015 caused some leaking in KLIA 2 airport which was complained by Tony Fernandes, the founder of AirAsia airline. Topic 3, dedicated mainly on this event. Corporate Social Responsibility (CSR) policies in Malaysia was discussed as Topic 4. And finally, Topic 5 was titles by domain experts as Korean foods talked about in social media in studied duration.

TABLE II. TOP 5 DISCOVERED TOPIC PROBABILITIES IN STUDIED DAYS

Days	T1	T2	T3	T4	T5
1	0.239693	0.016771	0.006988	0.010482	0.02376
2	0.119321	0.002983	0.855782	0.00195	0.001836
3	0.243409	0.011774	0.131815	0.002304	0.00256
4	0.216134	0.005175	0.028919	0.218874	0.007915
5	0.261624	0.013949	0.00403	0.645071	0.00775
6	0.313418	0.006366	0.032811	0.113124	0.005877
7	0.226906	0.618402	0.006598	0.015396	0.002933
8	0.193096	0.019646	0.003368	0.744878	0.003929
9	0.228458	0.014456	0.002834	0.142574	0.010204
10	0.204915	0.004998	0.011245	0.052062	0.007913
11	0.202906	0.010898	0.021277	0.006746	0.008303
12	0.244359	0.013002	0.008031	0.007648	0.008413
13	0.384447	0.021906	0.009858	0.012596	0.005476
14	0.292142	0.055133	0.008872	0.006971	0.012041
15	0.440781	0.016824	0.013459	0.01817	0.010767
16	0.347452	0.601274	0.003503	0.007962	0.003503
17	0.216173	0.676218	0.006367	0.005412	0.012416
18	0.26679	0.240852	0.006021	0.005558	0.008337
19	0.37839	0.036745	0.003937	0.003937	0.009186
20	0.169085	0.034069	0.008202	0.015142	0.011987
21	0.190219	0.013092	0.009241	0.005776	0.715056
22	0.184996	0.006284	0.005892	0.005499	0.754517
23	0.390703	0.013485	0.003194	0.005323	0.002839
24	0.786416	0.029789	0.006851	0.003277	0.113196
25	0.309557	0.018698	0.00554	0.024238	0.004848
26	0.270923	0.011217	0.005177	0.326143	0.00719
27	0.173396	0.010642	0.004069	0.011268	0.005947
28	0.375577	0.002539	0.001616	0.003693	0.00277
29	0.331361	0.005569	0.007309	0.011138	0.008354
30	0.136067	0.00421	0.781233	0.002852	0.003531

Profanities of all detected topics over days are presented in Table 2. In this table, the highest probability of every topic in each day are illustrated using gradient of red color in which highest probabilities are in darker red. Making it simpler to interpret, only top 5 topics are listed in Table 2 and other general ones were hided. For example, it is clear that topic 1, moving AirAsia flights from Terminal 2 to Terminal 1 at KKIA, is published almost in every day of Nov 2015 while Topic 5 was only discussed in 21st and 22nd . Incredibly, Paris attacks was detected correctly as the hot topic in social media for 16th and 17th Nov, just two days after this event happened in real world.

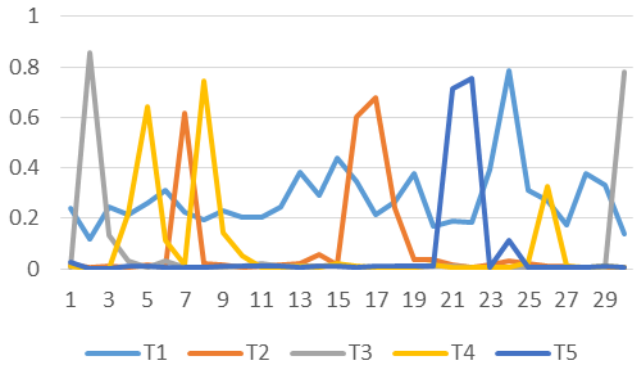

Fig.1. Dynamics of Topic over Days

Figure 1 depicts the dynamics of elicited social media topics in domain of Aviation and Airport Management over days during November 2015. As shown in this figure, Topic 1 (light blue) is the dominating topic for studied duration. The other ones have fluctuated influenced by real-world events. As a sample, Topic 2 (security and Paris attack) was peaked at 16th and 17th Nov just 2 days after the events happened on the evening of 13 November in Paris. Although, all these observations shows the success of LDA topic modeling algorithm developed in this research work, but we cannot state that it works quite perfectly as this claim is far from reality in all machine learning methods.

V. DISCUSSION AND CONCLUTION

In this study, we developed the LDA method to detect top topics in social media content for the domain of Aviation and Airport Management. Utilizing the developed topic model, we also visualized the dynamics of topics over days based on a probabilistic approach.

The performance of the LDA topic modeling algorithm developed in this research work was evaluated using a qualitative method. To this end, we asked domain expert to investigate the detected topics along with assigned keywords and compare the results with their own interpretation about the top topics of studied datasets. The experts' investigation results revealed that the developed LDA algorithm, successfully detected the main social media topics in studied domain. In

addition, the dynamics of topics made it clear that topic 1, moving AirAsia flights from Terminal 2 to Terminal 1 at KKIA, has been the dominating topic in November 2015, while the other topics were discussed more and less in social media affecting by the real world events.

The developed topic model can be used in different domains and applications such as finding scientific topics from articles published in conference proceedings and journals, detecting main customer request themes from CRM log files, and preventing terrorism attacks by mining related topics in social media content and digital conversations. For the future works, applying the developed LDA method to every single records published by authors in social media and then aggregating the detected topics for every day could be a useful extension of this study.

ACKNOWLEDGMENT

This research work is supported by Berkshire Media (formerly known as Berkshire Consulting &Management Services Sdn Bhd (974773-H)).

REFERENCES

1. Ravi, K. and V. Ravi, *A survey on opinion mining and sentiment analysis: tasks, approaches and applications.* Knowledge-Based Systems, 2015. **89**: p. 14-46.
2. Rohani, V.A. and O.S. Hock, *On social network web sites: definition, features, architectures and analysis tools.* Journal of Computer Engineering, 2009. **1**: p. 3-11.
3. Rohani, V.A. and S. Shayaa, *Utilizing Machine Learning in Sentiment Analysis: SentiRobo Approach*, in *2nd International Symposium on Technology Management and Emerging Technologies*. 2015: Langkawi, Malaysia.
4. Hu, Y., et al., *Interactive topic modeling.* Machine learning, 2014. **95**(3): p. 423-469.
5. Hong, L. and B.D. Davison. *Empirical study of topic modeling in twitter*. in *Proceedings of the First Workshop on Social Media Analytics*. 2010. ACM.
6. Arora, S., et al., *A practical algorithm for topic modeling with provable guarantees.* arXiv preprint arXiv:1212.4777, 2012.
7. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* the Journal of machine Learning research, 2003. **3**: p. 993-1022.
8. Tang, J., et al. *Understanding the limiting factors of topic modeling via posterior contraction analysis*. in *Proceedings of The 31st International Conference on Machine Learning*. 2014.
9. Ramage, D., S.T. Dumais, and D.J. Liebling, *Characterizing Microblogs with Topic Models.* ICWSM, 2010. **10**: p. 1-1.
10. Wang, C. and D.M. Blei. *Collaborative topic modeling for recommending scientific articles*. in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011. ACM.
11. Ling, X., et al., *Generating gene summaries from biomedical literature: A study of semi-structured summarization.* Information Processing & Management, 2007. **43**(6): p. 1777-1791.
12. Hu, Y., et al., *A linked-data-driven and semantically-enabled journal portal for scientometrics*, in *The Semantic Web–ISWC 2013*. 2013, Springer. p. 114-129.
13. Ramage, D., et al. *Topic modeling for the social sciences*. in *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*. 2009.
14. Grimmer, J., *A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases.* Political Analysis, 2010. **18**(1): p. 1-35.
15. Nguyen, T.H. and K. Shirai. *Topic modeling based sentiment analysis on social media for stock market prediction*. in *Proceedings of the 53rd Annural Meeting of the Association for Computational Linguistics*. 2015.
16. Mimno, D., *Computational historiography: Data mining in a century of classics journals.* Journal on Computing and Cultural Heritage (JOCCH), 2012. **5**(1): p. 3.
17. Deerwester, S.C., S.T. Dumais, and R.A. Harshman, *Indexing by latent semantic analysis.* 1990.
18. Papadimitriou, C.H., et al. *Latent semantic indexing: A probabilistic analysis*. in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. 1998. ACM.
19. Hofmann, T. *Probabilistic latent semantic indexing*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999. ACM.
20. Mcauliffe, J.D. and D.M. Blei. *Supervised topic models*. in *Advances in neural information processing systems*. 2008.
21. McCallum, A., X. Wang, and N. Mohanty, *Joint group and topic discovery from relations and text*. 2007: Springer.
22. Zhang, H., et al. *Probabilistic community discovery using hierarchical latent gaussian mixture model*. in *AAAI*. 2007.
23. Nallapati, R.M., et al. *Joint latent topic models for text and citations*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. ACM.
24. Chang, J., J. Boyd-Graber, and D.M. Blei. *Connections between the lines: augmenting social networks with text*. in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009. ACM.

25. Liu, Y., A. Niculescu-Mizil, and W. Gryc. *Topic-link LDA: joint models of topic and author community*. in *proceedings of the 26th annual international conference on machine learning*. 2009. ACM.
26. Li, D., et al. *Community-based topic modeling for social tagging*. in *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010. ACM.
27. Zhai, K., et al. *Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce*. in *Proceedings of the 21st international conference on World Wide Web*. 2012. ACM.
28. Daud, A., *Using time topic modeling for semantics-based dynamic research interest finding*. Knowledge-Based Systems, 2012. **26**: p. 154-163.
29. Vavliakis, K.N., A.L. Symeonidis, and P.A. Mitkas, *Event identification in web social media through named entity recognition and topic modeling.* Data & Knowledge Engineering, 2013. **88**: p. 1-24.
30. Lim, K.W., C. Chen, and W. Buntine. *Twitter-network topic model: A full Bayesian treatment for social network and text modeling*. in *NIPS Topic Model workshop*. 2013.
31. Nguyen, M., T. Ho, and P. Do. *Social networks analysis based on topic modeling*. in *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*. 2013. IEEE.
32. Yang, S.-H., et al. *Large-scale high-precision topic modeling on twitter*. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014. ACM.
33. Ostrowski, D.A. *Using latent dirichlet allocation for topic modelling in twitter*. in *Semantic Computing (ICSC), 2015 IEEE International Conference on*. 2015. IEEE.
34. Team, R.C., *R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013.* Document freely available on the internet at: http://www. r-project. org, 2015.