Using Topic Modelling to determine the relationships between Hashtags and Tweets.

Olashile Adebimpe

Abstract

Various unsupervised topic modeling approaches have been employed to the Twitter dataset but have been faced with drawbacks caused by the sparsity of the short word length of tweets. So we proposed a semi-supervised approached of modeling topics by employing tweets aggregation and priors to determine the relationship between the modeled topic and the underlying related hashtag of the tweets. Our study uses two evaluation measures; Topic Coherence Analysis and our NLTK Similarity Measure Algorithm, to evaluate the result from the topic models and also compare the relationship with its related hashtag respectively. Our obtained results show that GuidedLDA shows more insight and relationship value for modeled topics that have some high coherence similarity with the hashtag.

1. Introduction

Topic Modelling in Natural language processing is a kind of probabilistic generative model and have been very useful and applicable in so many fields of computer science, with much focus based on text mining. Since the model was first proposed, it has received a lot of attention and gained interest among researchers in so many domains outside of computer science. Apart from text mining, topic modeling has been used successfully to solve various domain challenges in computer vision [1][2], population genetics, social networks [3]and also in Bioinformatics[4]. Researchers in the medical and biological domain have continuously employed topic modeling for biomedical text mining and clinical informatics task, due to its superiority in the analysis of large-scale document collections to yield better result[4].

In recent years, topic modeling has also been employed on user-generated content and information created in form of text, video, images, emoticons posted on social media platforms such as Facebook and Twitter[5]. These unstructured form of content published on all these platforms in real time has helped provide more information containing a diverse opinion and topics from over millions of social media

users.The information generated on these different social media platform provided timely, actionable and sometimes fact base insight about social topics.

There are various approaches employed for discovering the hidden patterns in large unstructured text and the two popular approaches are the Probabilistic approaches which assume every document is a mixture of a smaller number of topics e.g. Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA) and the Matrix factorization approaches such as Non-negative Matrix Factorisation (NMF) which employes linear algebra to help decompose a matrix. However, being major approaches in topic modeling, they still have some minor scenarios where modeling went wrong by not identifying the topic in accordance to the underlying structure of the document as these approaches focus on document level co-occurrence information to group semantically related words into a single topic.

In this paper, we planned to create a system aimed at finding correlation and relationship between social media content and its assigned topic by applying a semi-supervised approached to topic modeling algorithm and using some similarity measure algorithm to measure the similarity between the topic of the content and the series of modeled topics. Our choice of social media platform will be Twitter due to its abundance of content from the enormous amount of tweets posted every second. However, detecting topics in tweets can be a challenging task due to their informal type of language and since tweets usually are more incoherent than traditional documents[6]. The rest of the text is outlined as follows: section 2 describes the topic modeling task and some previous work in the field, section 3 describes the sampling and preprocessing methods we employed, while section 4 describes our unsupervised approach to topic modeling, section 5 details how the set of experiments were carried out and section 6 discusses sums up the results and point to some direction for future research.

2. Topic Modelling

Topic Modelling are statistical methods developed from new techniques used for finding patterns of words in a document collection using hierarchical probabilistic models. It provides a convenient way to analyze a large set of unclassified text. Topic models rely strongly on bag-of-words assumption which is ignoring the information from the ordering of words. According to Seungil and Stephen, 2010, “Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, a corresponding low-rank representation of

the high-dimensional histogram can be obtained for each document" [7]. In the early days of Topic Modelling, Latent semantics Analysis (LSA) - a dimensionality reduction method was introduced to mine topics by constructing term matrix with the help of singular value decomposition. However, the major drawback of this topic model is the inability for LSA to capture polysemy in the corpus given and also the amount the amount of time it takes to discover topics and hidden word patterns in a given corpus [9]. In recent times, Latent Dirichlet Allocation (LDA) which assumes that documents are represented as a mixture of latent topics, where each topic is characterized by distribution over words, was proposed by Blei & Jordan in 2003 [8] to address the major drawback of Latent semantics Analysis (LSA) and also LDA has become the standard tool in topic modelling [9][10] and a number of variants and extensions has also been proposed. There has been a substantial amount of work that has been done leveraging on LDA and its variants. Hong et. al.[14] proposed the use of LDA to derive a topic model within a microblogging (Twitter) environment by training topic model on aggregated measures which resulted in a topic mixture distribution, thereby supporting a good set of supplementary features in classification problem. Zhao et. al. [11] applied a Twitter variant model which he named Twitter-LDA model to discover topics by allowed for a comparison to traditional news media allowing for input into data mining applications. Biro et.al. [12] applied LDA to web spam filtering also Endres et. al. [13]extended LDA model to support feature generation and hypothesis generation in 3D range data. Also, Lau et al. [14] modify the LDA by adding contribution factor parameter which helps LDA to have a set of constantly evolved topics and introducing the concept of time slices which was used to partition the document. Song et al. [15] utilize Dirichlet-Multinomial Regression (DMR) for extracting tends on Twitter data. DMR which is also an extension of LDA, that allows conditioning on arbitrary document feature by including long-linear prior to document-topic distribution. The above approaches are based on variations LDA prior which changes unsupervised characteristic of LDA to a semi-supervised approach

3. Sampling and Preprocessing

In order to access Twitter dataset, we used Tweepy, a python library for accessing Twitter for the collection of tweets. We sampled tweets and hashtags from some selected hashtag during the period of this research and we collected an approximately 330k tweets related to 40 hashtags and will be discussing the text analysis preprocessing steps we needed to carry out on the sample tweets.

Our approach can be majorly into 3 main categories.

I. The Preprocessing Stage

Our preprocessing stage consists of several parts which include the removal of Uniform Resource Locator(URL), emojis, punctuation, and stop words. Duplicated tweets were also removed as the extraction of tweets was carried out serval time for a single hashtag. Cleaned each tweet by removing the "RT" which means Retweet, removed all usernames and hashtags in the tweets as they both begin with @ and # respectively. We applied case conversion and lemmatization, which is removing the inflection endings only and returning the base form of the word called lemma.

II. The sentiment analysis and Tweet aggregation.

After preprocessing of tweets, we probably left with just a few tokens in the tweet as most have been filtered off by the preprocessing task above and this poses a serious difficulty to modeling tweets because of their sparseness as a small number of tokens might not contain enough or sufficient data to establish satisfactory token co-occurrence. Therefore we applied some sentiment analysis to group related tweet token with the same sentiment about the hashtag and also extracting the period of the day the tweet was posted on Twitter. The combination of these two approaches was employed to create some tweet aggregation that shares the same sentiment and also posted almost same period of the day.

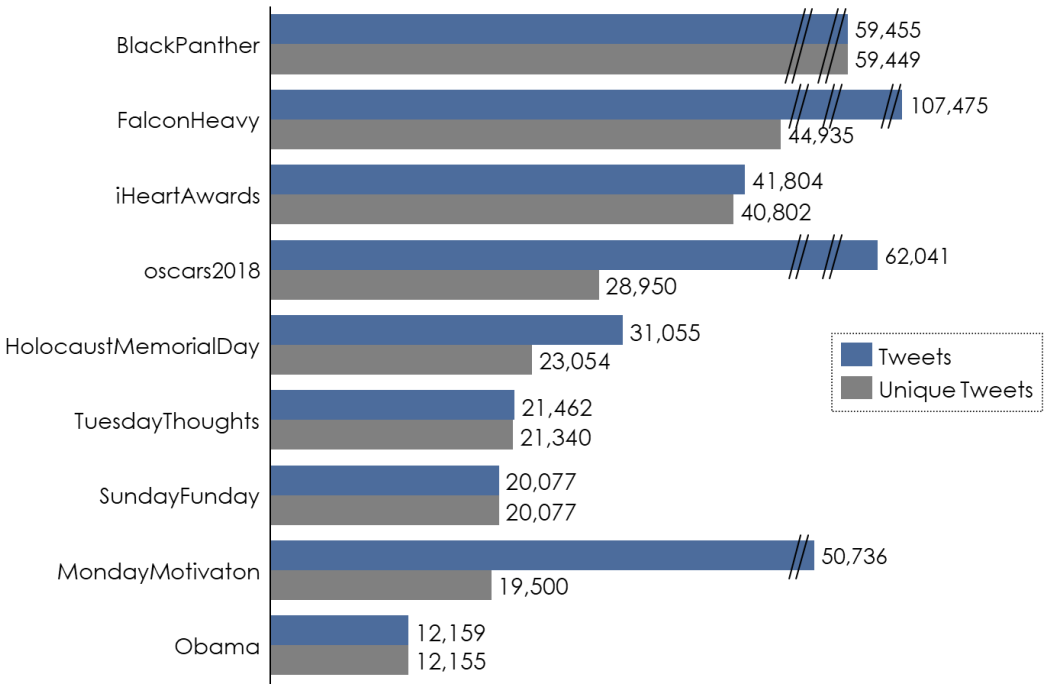

Figure 1: Distribution of sampled Tweets

III. Tokenisation and Document-Term Matrix Creation.

We employed the popular NLTK library to tokenize our aggregated tweets and passed them for further preprocessing by applying CountVectorizer class with appropriate parameters. We used the CountVectorizer instead of TFIFDVectorizer as LDA does not apply on the sparse matrix and converting a sparse matrix to a count matrix might just result in so many zero matrix elements which will affect negatively out topic models.

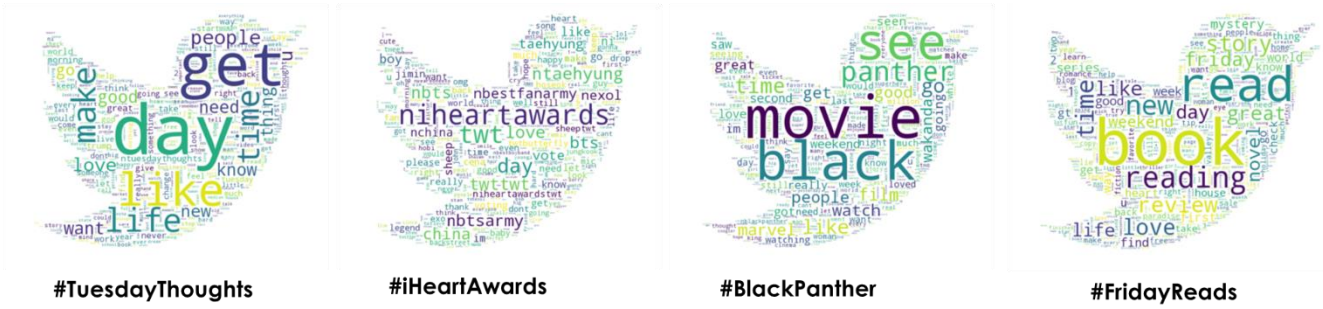

Figure 2: Sample Preprocessed Tweet Tokens

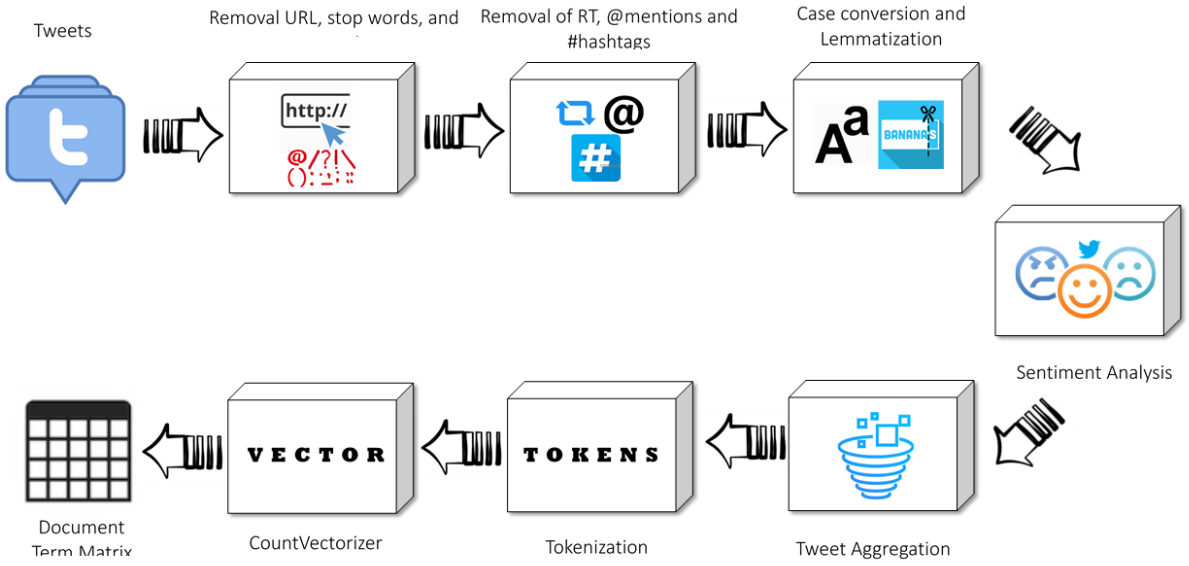

Figure 3: Preprocessing Steps

4. Topic Modelling Approach

Various Topic modeling techniques have already been proposed and employed for different methodology and content of data, but there still remains two broad categorizations of the various techniques which are the

I. The Probabilistic Approach that assumes document viewed as a mixture of a small number of topics and the word and documents get probability score for each topic. An example is Latent Dirichlet Allocation (LDA) [8] and
II. The Matrix Factorisation approaches that apply methods from linear algebra to decompose a single matrix into a set of smaller matrices. An example is Non-negative Matrix Factorisation (NMF)[16]

However, these algorithms are unsupervised classes of machine learning algorithm and they are generally good at grouping words into topics which always results in topics which are neither entirely meaningful not very effective in extrinsic task employed for [17] because they implicitly use document level co-occurrence information to group semantically related words into a single topic.

So we propose to employ and use the simple and effective method of adding priors to the model to guide the direction of the modeled topics[18]. We used a semi-supervised model name GuidedLDA or SeededLDA which was created by Allen Riddell and Tim Hopper. GuidedLDA was developed and implemented on the Latent Dirichlet allocation (LDA) using collapsed Gibbs sampling by allowing the certain setting to made to influence the direction in which the topic converges. These settings include priors which are called seed words are provided to model and which contains a representative of the underlying topics in the corpus and the confidence intervals of the prior provided to the model.With the seed list provided to the model, out model is only encouraged to follow the direction of the seed list when trying to converge on a topic and if the document has a more compelling reason not to follow the seed list, the model the topic in a ways ignoring the prior provided into the document.[19]

$$
\begin{aligned}
&for\ document\ in\ all_{tweets}:\\
&\quad for\ word\ in\ all_{word}:\\
&\quad\quad for\ topic\ in\ all_{topic}:\\
&\quad\quad\quad word_{in_{topic}} = count(all\ document\ that\ belongs\ to\ a\ topic)\\
&\quad\quad\quad document_{in_{topic}} = count(words\ in\ document\ that\ belong\ to\ a\ topic)\\
&\quad\quad\quad token_{in_{topic}} = count(all\ assignment\ in\ topic)
\end{aligned}
$$

$$p(topic|word, document) = \frac{word_{in_{topic}} * document_{in_{topic}}}{token_{in_{topic}}}$$

$$word_{in_{topic}} = \max(p(topic|word\ , document))$$

Seed Extraction

Considering the nature of our dataset - microblogging site, generating a seed list might be too complicated due to diverse nature of the corpus, so we made few assumptions in order to come up with the right approach to create a seed list for our topic model. We assumed that major tokens in topics or subject or title of any document always belong to the Noun part of speech and all other tokens in the title of the document are always telling us more information about the head of the sentence – Noun. So we used Part-of-speech (POS) taggers to extract all correctly spelled token tagged as a noun to create our seed list. Therefore making $document_{in_{topic}} = count(words\ in\ document\ that\ belong\ to\ a\ topic)$ to be higher for seeded words in topic which implies that the probaility of a topic being around the seeded list with be higher.

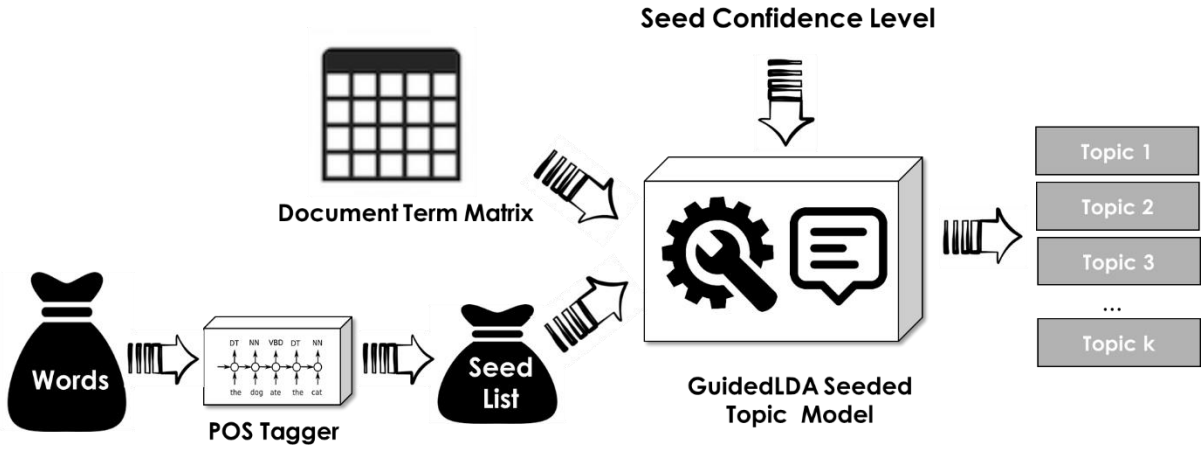

Figure 4: Our Topic Modeling Approach

5. Topic Modeling Experiments

In this section, we would be explaining the major parameter selection that was made for the model, the selection from our sampled preprocessed dataset, and result comparison with a standard LDA topic model using the topic coherence analysis measure and the NLTK similarity measure.

I. Parameter Selection: The key required parameter which is the number of topics to be extracted from the corpus was set to 10, likewise the number topic terms that compose a single topic, also the number iteration our model should make before converging was set to the maximum.

II. Dataset Selection: Due to the time it takes our model to converge on a topic, we had to sample different category of our dataset used for the experiment. We further categorized out dataset into 4 main categories which are hashtags on popular Individuals, events around the world, some hashtags related to days and some random hashtags. Below is the brief break down of the selected category.

Hashtags			
Individuals	**Events**	**Days**	**Random**
Obama	SuperBowl	TuesdayThoughts	PressforProgress
Trump	iHeartAwards	MondayMotivation	BreakingNews
	FalconHeavy	ThursdayThoughts	
	HolocaustMemorialDay	FridayReads	

Figure 5: Table showing selected hashtag from 4 different category

III. Evaluation Techniques:

- Topic Coherence which gives the extent to which the top terms representing our modeled topic are semantically related relative to the background corpus
- Our Similarity measure approach was built on NLTK wordnet.sysnets to measure the similarity between the modeled topic and the related hashtag.

6. Results and Observations

Ten topics consisting of 10 terms were generated for each of the models and below is a word cloud representation of the modeled topic using our approach and the standard LDA. And from figure 6 below, we could see kind of words produced by the two different models. The GuidedLDA model provided more words with high-frequency using the seeded list provided and also the terms in the modeled topic has high coherence score with the underlying semantic structure of the corpus.

The similarity measure compared each term in the topic model with the attached hashtag using the NLTK wordnet.sysnets and attached values between 0.0 and 1.0 representing at least similar to most similar.

- For #Trump, #HolocaustMemorialDay #ThursdayThoughts, GuidedLDA produce more terms which have more semantics similarity with the background corpus than LDA while the otherwise occurred when tested on hashtag #PressforProgress

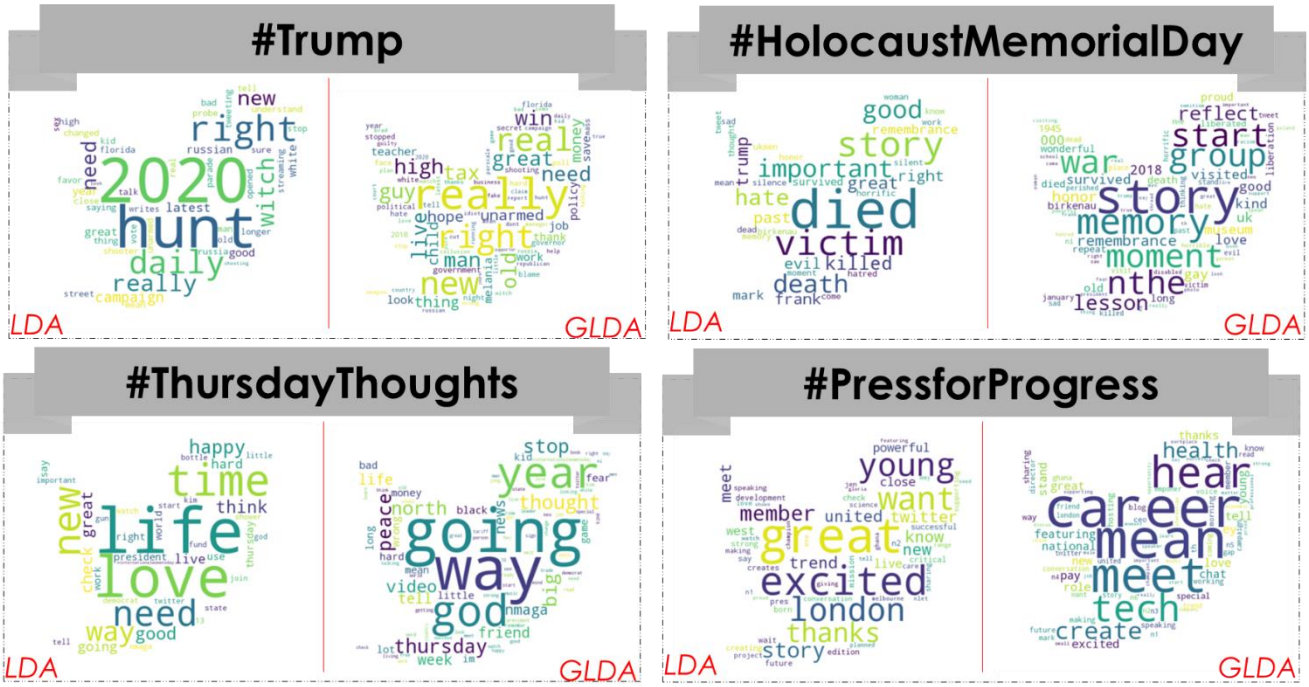

Figure 6: Word Cloud Representation of Modelled Topics

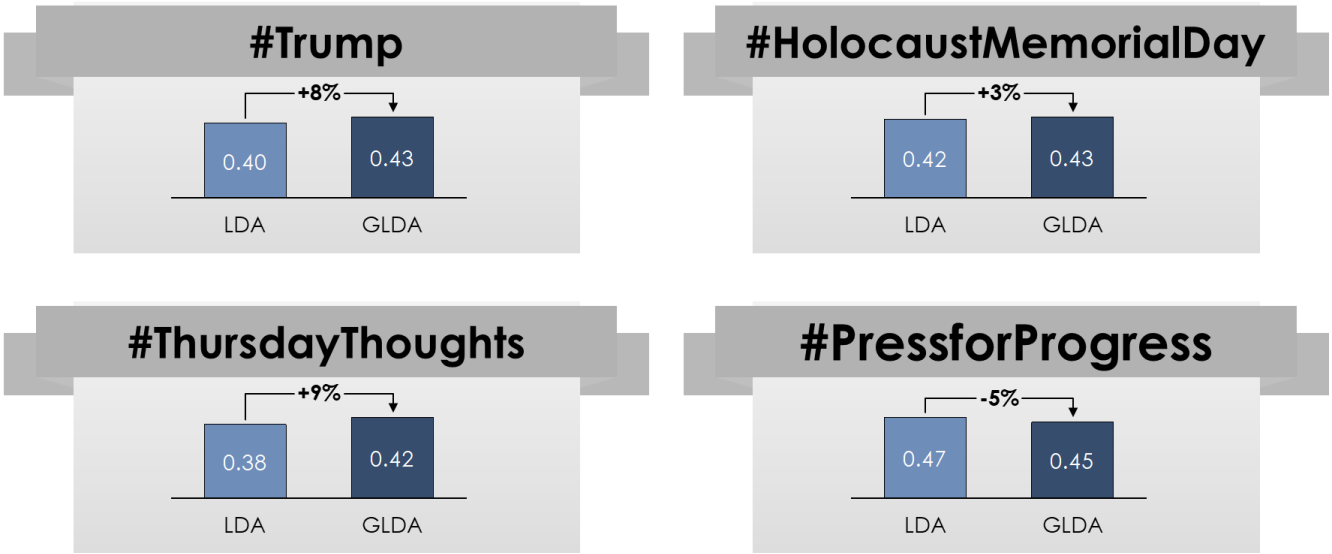

Figure 7: Topic Coherence Analysis Measure

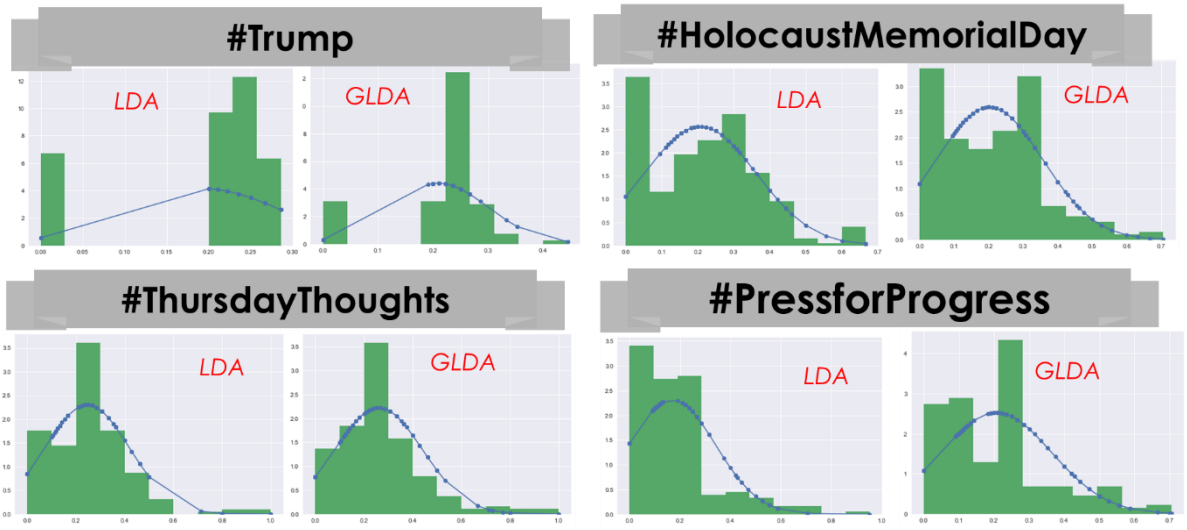

Figure 8: Normal Distribution of the Similarity Measure between the Modeled Topic and Hashtag

7. Future Works

In this paper, we proposed a semi-supervised approach for modeling topics using aggregation method and seed words. The advantage of our approach over the unsupervised approach is to provide guidance to the model on which topic or terms it should converge around.However, the obtained result could still be enhanced by improving on the key parameter selection for our model and also to optimize and improve the algorithm used for seed list generation.

We also propose this approach tried out on more structured topic modeling problem which employes standard information retrieval evaluation techniques to evaluate the result of the model.

References

[1] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 524–531, 2005.

[2] W. Luo, B. Stenger, X. Zhao, and T.-K. Kim, "Automatic Topic Discovery for Multi-Object Tracking," in *Proceedings of the Twenty-Ninth AAAI COnference on Artificial Intelligence*, 2015, no. 1, pp. 3820–3826.

[3] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized POI recommendations," *IEEE Trans. Multimed.*, vol. 17, no. 6, pp. 907–918, 2015.

[4] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *Springerplus*, vol. 5, no. 1, 2016.

[5] K. Patel, O. Hoeber, and H. J. Hamilton, "Real-Time Sentiment-Based Anomaly Detection in Twitter Data Streams," *Adv. Artif. Intell.*, vol. 9091, pp. 196–203, 2015.

[6] A. O. Steinskog, J. F. Therkelsen, and B. Gambäck, "Twitter Topic Modeling by Tweet Aggregation," *Proc. 21st Nord. Conf. Comput. Linguist.*, no. May, pp. 77–86, 2017.

[7] T. Hofmann, "Unsupervised learning by probabilistic Latent Semantic Analysis," *Mach. Learn.*, vol. 42, no. 1–2, pp. 177–196, 2001.

[8] D. M. Blei, B. B. Edu, A. Y. Ng, A. S. Edu, M. I. Jordan, and J. B. Edu, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[9] G. Maskeri, S. Sarkar, and K. Heafield, "Mining business topics in source code using latent dirichlet allocation," *Proc. 1st Conf. India Softw. Eng. Conf. - ISEC '08*, p. 113, 2008.

[10] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, 2010, pp. 80–88.

[11] W. X. Zhao *et al.*, "Comparing Twitter and Traditional Media using Topic Models," *Proc. 33rd Eur. Conf. Adv. Inf. Retr.*, pp. 338–349, 2011.

[12] I. Bíró, J. Szabó, and A. a Benczúr, "Latent dirichlet allocation in web spam filtering," in *Proceedings of the 4th international workshop on Adversarial information retrieval on the web AIRWeb 08*, 2008, p. 29.

[13] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard, "Unsupervised Discovery of Object Classes from Range Data using Latent Dirichlet Allocation," *Proc. Robot. Sci. Syst.*, 2009.

[14] J. Lau, N. Collier, and T. Baldwin, "On-line Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online," *Int. Conf. Comput. Linguist.*, vol. 2, no. December, pp. 1519–1534, 2012.

[15] M. Song and M. C. Kim, "RT2M : Real-time twitter trend mining system," in *Proceedings - 2013 International Conference on Social Intelligence and Technology, SOCIETY 2013*, 2013, pp. 64–71.

[16] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization,"

Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[17] J. Chang, S. Gerrish, C. Wang, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” *Adv. Neural Inf. Process. Syst. 22*, pp. 288--296, 2009.

[18] J. Jagarlamudi and H. Daum, “Incorporating Lexical Priors into Topic Models,” *Umiacsumdedu*, pp. 204–213, 2009.

[19] “How we Changed Unsupervised LDA to Semi-Supervised GuidedLDA.” [Online]. Available: https://medium.freecodecamp.org/how-we-changed-unsupervised-lda-to-semi-supervised-guidedlda-e36a95f3a164. [Accessed: 16-Apr-2018].